Using CNNs to Explore Painting Themes

Christina Hung[1], Nick Troccoli[1], Mana Lewis[2]*
Stanford University
[1]{chung888, troccoli}@stanford.edu
[2]mana@chezmana.com

Abstract

Image recognition is a task that has been largely tackled in the past few years using various convolutional neural network (CNN) architectures. However, paintings are a substantial subset of image data that are slightly more difficult to analyze due to the sometimes abstract nature of paintings and the artistic interpretations made to create them, across artists and painting time periods. To explore possible underlying characteristics of certain paintings and themes, we focus on thematic content of paintings across artists and time periods; specifically, adult portraits, and whether we can successfully classify paintings according to this theme. We use a dataset of over 35K paintings and build on top of VGGNet-16, a successful competitor in the ImageNet challenge, with additional featurizers such as color histograms and Gram matrices. We achieve over 80% classification accuracy with our best model.

1. Introduction

From organizing art collections, to auto-categorizing and curating images on Flickr and Google Images, to clustering art on organization sites such as Pinterest, detecting thematic similarity in paintings is a significant yet challenging computer vision task. By thematic similarity, we are referring to themes such as seas and oceans, mountains, male portraits, animals, and other high level categorizations. However, due to the sometimes abstract nature of paintings and the artistic interpretations made to create them, thematic comparison is trickier for paintings than normal images, which are complete representations of the real world. But uncovering signature characteristics of themes across artists and time periods makes this a rewarding challenge. We focus on one specific theme, portraits, and how to classify between paintings of portraits vs. other themes.

For our model, we chose to use Convolutional Neural Networks (CNNs), which have given a useful architecture for large-scale image and video recognition [9] [16] [13] [14]. The ImageNet Large-Scale Visual Recognition Challenge[9], where models are tasked with classifying images into 1000 categories, has played an important role in advancing deep visual recognition architectures, particularly CNNs. Because of this, and a curiosity to see whether ImageNet success carries over to success classifying paintings, we decided to build on top of a past ImageNet winner, VGGNet-16 [14], which contains a combination of convolutional and fully-connected layers. We add additional layers at the beginning to allow inputting additional potential indicators of theme including color histograms and Gram matrices. At a high level, our model takes as input a single painting and outputs a score of 1 if it predicts the painting is a portrait, and 0 if not. Overall, our results show that both color and covariance play a small yet significant role in determining whether a painting is a portrait, regardless of time period and artist.

2. Related Work

Digitized art paintings have been available for many years, and artwork classification is a problem that has been often explored. While the thematic content of art is oftentimes classified by hand, many learning algorithms have been used in context of art stylistic classification [4]. To this end, several input image featurizers are often used, usually as a fusion of features [4], [12] to improve model robustness. In [17] and [4], edge texture detection, steerable filter decomposition and color histogram features were used to improve classification results. Of all these features, using a color histogram stands out the most in terms of thematic painting detection.

This is because colors appear to be the most notable feature that ties together theme similarity across paintings– that is, paintings with similar themes more likely than not have the same color scheme: for instance, paintings with the theme ”seas-and-oceans” are likely to have a

*Mana generously provided our dataset and is not enrolled in CS231N.

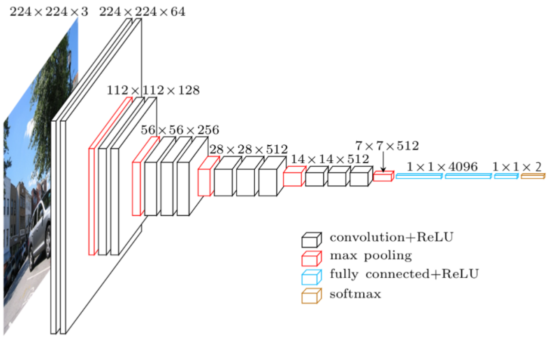

Figure 1. VGGNet-16 architecture. [2]

predominantly blue backdrop, while those with the theme ”pastoral” will have a predominantly green background. Moreover, paintings with the same theme may have similar distributions of light and dark colors. [17] states that the most sophisticated approach for measuring color similarity is based on the dominant colors (red, green, and blue since the human visual system cannot perceive more than these three colors), and their respective percentages. This has been used successfully for edge retrieval and segmentation (and is therefore useful in detecting objects in images) [6], [11], [10].

As successful as the experiments in [17] were, only a dataset of 353 paintings (with 5 genres) was used–what brought robustness to such a small dataset was the variety of handcrafted features that were specially tailored to the images in the small dataset. With our fairly large dataset, however, we were able to demonstrate robustness of our model simply with its size (around 15,010 paintings); we used only features such as color histograms and Gram matrices (for painting style) [8] that seemed appropriate in the context of artwork. We also explored the binary code representations of image features (known as ”Picture Codes” or PiCoDes) used in [7], [5], and [15], with a fusion of our feature vectors that we first extracted from each painting, then merged into the input image fed into the first layer of our classification model architecture, but ultimately did not have time to complete this implementation.

Additionally, as mentioned above, previous work in image classification achieved with CNNs has shown to be successful in [9], [16], [13], and [14]. Many learning methods used by previous art-classification-networks made use of a variety of ImageNet CNNs to achieve state-of-the-art performance; similarly, we used the VGGNet-16 architecture [14] for our classification model of painting themes.

3. Methods

As mentioned previously, our methods centered around reusing and building on the architecture of VGGNet-16[14], a former winner of the ImageNet challenge. VGG, as seen in Figure 1, takes as input a 224x224x3 RGB image, which is passed through a series of 3x3 and 1x1 filter convolutional layers, as well as max-pooling layers, which you can see in red interspersed among the convolutions. Following this are 3 fully-conneted layers that take the state 4096 channels down to just 2 - one for the class of ”adult portrait”, and one for the class of ”non-portrait”.

We use the same architecture as our baseline [1] because of its success in the ImageNet challenge, and thoughts that the potentially more photo-esque qualities of portrait paintings may lead to good results here as well. Moreover, we were curious if techniques applied to photographic images could in general be successfully applied to the more abstract painting.

Because VGG accepts a 224x224x3 RGB image as input, we first resize all input paintings to be 224x224x3 using Bilinear Interpolation. We decided to resize rather than crop because, we suspected that the portrait theme may be deducible only from a specific portion of a painting, which may be lost when cropping.

On top of VGG, we also add 2 additional featurizers that input to the first layer:

1. A 3x16 per-channel **color histogram** for each painting
2. A 32x32 **Gram covariance matrix** for each painting

The color histogram[2] is a 3x16 tensor where each row contains the normalized frequencies of R, G and B values, respectively, in that painting among 16 buckets (0-15, 16-31, etc.). For instance, Figure 2 displays one slice of this histogram tensor, a histogram of red values for the image displayed in Figure 4.

To feed this into our VGGNet architecture, we flatten the 3x16 histogram to be 1x48, pass it through a fully-connected layer with 224 hidden units, and then resize the result to 224x1x1 which we add to the 224x224x3 image input. We then input the resulting 224x224x3 tensor into VGG. We decided on this architecture so that the model could learn how much weight to give to color histograms, and also so that the color histogram data could be present in all parts of the resulting tensor.

[1] Starter code from https://gist.github.com/omoindrot/dedc857cdc0e680dfb1be99762990c9c

[2] Code based on https://stackoverflow.com/questions/34130902/create-color-histogram-of-an-image-using-tensorflow

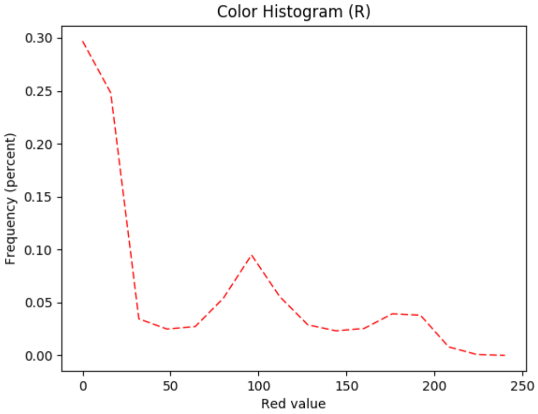

Figure 2. Distribution of red values in Figure 4

The Gram matrix [3] is a 32x32 matrix that represents covariance in pixels throughout the painting. The Gram Matrix in general is a CxC matrix computed from a 1xHxWxC image. In order to increase the amount of data stored in our gram matrix (which would be 3x3 for our input image), we first pass the 224x224x3 input image through a convolutional layer with 32 7x7 filters, which results in a 224x224x32 tensor. We then calculate the 32x32 Gram matrix of this output. Mathemtically, we resize the input tensor into a 1xCxM feature map F, for which the 1xCxC Gram matrix is represented as

$$G_{ij} = \sum_k F_{ik} F_{jk} \tag{1}$$

To feed this into our VGGNet architecture, similar to our approach with our color histogram, we flatten the 32x32 matrix to be 1x1024, pass it through a fully-connected layer with 224 hidden units, and then resize the result to 224x1x1 which we add to the 224x224x3 image input. We then input the resulting 224x224x3 tensor into VGG. Similar to our approach with histograms, we designed this architecture so that the model could learn how much weight to give to the Gram matrices, and also so that the covariance data could be present in all parts of the resulting tensor.

With VGG and these featurizers, we implemented 4 model versions:

1. **SimpleModel**: our baseline model that is just the VGG architecture
2. **HistogramModel**: our baseline model with the additional layer mentioned above for color histogram input
3. **GramModel**: our baseline model with the additional layer mentioned above for Gram matrix input
4. **JointModel**: our baseline model combined with both color histogram and gram matrix input.

Of note is our Joint Model, which generates a color histogram with a fully-connected layer as mentioned above, a Gram matrix with a conv layer and fully-connected layer as mentioned above, concatenates them together, and then *concatenates the result* to the input image. This ensemble allows our model to selectively weight the importance of each feature in determining whether a painting is an adult portrait.

Altogether, we compared the performance of these different models to see if we could deduce what features, if any, were more indicative of adult portrait paintings. In order to do this, however, we first needed to collect and organize a dataset of paintings.

4. Dataset

Our dataset is drawn from the WikiArt Dataset [3] (around 35,750 images), a visual art encyclopedia of searchable works of fine art, which was collected by the Stanford Logic Group. The collection as provided describes 35,749 unique digitized paintings of varying sizes. Each painting has been labeled with the following metadata: artist name, thematic content, painting title, art movement (style), and genre. The sparseness of this dataset (see Figure 3) Led us to narrow in on a specific theme, adult portraits, which is a combination of the two most popular themes, 'male-portraits' and 'female-portraits'. Figures 4 and 5 are sample paintings from our dataset.

Our preprocessing consisted of extracting all valid male-portrait and female-portrait paintings (a total of 7505 paintings). Next, we randomly selected 300 images (discarding themes with less than that) from each remaining category so as to ensure treatment of each alternate theme equally; this resulted in 15010 total samples, 50% adult portrait, 50% non portrait, which were split into $\frac{3}{5}$ train (9006 paintings), $\frac{1}{5}$ val (3002 paintings), and $\frac{1}{5}$ test (3002 paintings), based on the ratios used in CIFAR-10 [1].

5. Experiments/Results/Discussion

Our resulting Joint model, which consists of the 224x224x3 painting, color histogram, and Gram matrix as inputs, achieved 80.3131% accuracy on our test set. In order to get here, we ran various experiments to determine optimal hyperparameters, as well as optimal featurizers.

5.1. Hyperparameters

We did thorough hyperparameter searches for both the learning rate and the dropout keep probability to maximize

[3] Taken from CS 231N assignment 3

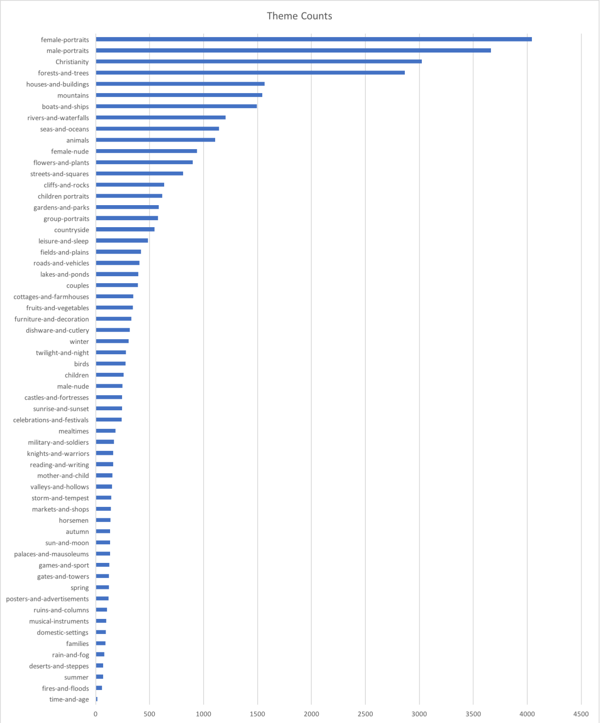

Figure 3. Thematic content distribution of the WikiArt dataset

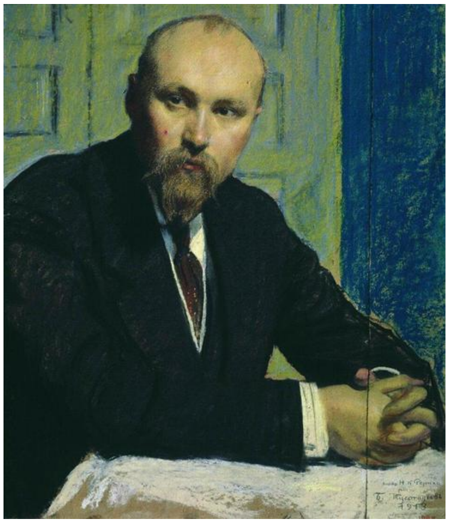

Figure 4. Example male-portrait-themed painting [3]

our overall performance. As shown below in Figure 6, we found the optimal learning rate for our model to be 0.005.

For dropout probability, as shown below in Figure 7, we found the optimal dropout keep probability for our model to be 0.8.

This is expected, as our dataset size was small, and therefore we hypothesize that a low dropout rate was necessary for optimal learning.

Figure 5. Example seas-and-oceans-themed painting [3]

	0.001	0.005	0.01	0.05	0.1
TRAIN ACC	65.26	70.88	69.84	50.19	50.19
VAL ACC	66.56	73.15	70.55	48.87	48.87

Figure 6. Hyperparameter search for learning rate using our JointModel

	0.01	0.2	0.4	0.6	0.8	0.9
TRAIN ACC	49.81	59.55	66.68	71.35	73.75	71.32
VAL ACC	51.13	61.29	68.12	72.55	74.35	72.25

Figure 7. Hyperparameter search for dropout using our JointModel

MODEL	TEST ACC
SIMPLE	80.0799
HISTOGRAM	77.0486
GRAM	78.3811
JOINT	80.3131

Figure 8. Performance of each model over 10 epochs

5.2. Comparing Models

Overall, as we expected and shown in Figure 8, our joint model performed the best, with 80.3131% accuracy on the test set. Surprisingly, in second place was our baseline model, followed by our two individual featurizer models.

From this, we suspect that overrelying on one featurizer may skew our results worse than weighting multiple featurizers in our model which can be more evenly balanced.

5.3. Results and Analysis

Focusing in on our best model's results, we did in-depth analysis of the 80.3131% accuracy score, what our model did well, and what our model could improve. We generated a telling chart of the performance scores for our model, on a scale of 0 - 100, of various other themes of paintings in the dataset - in other words, how successful our model was at labeling each of these themes as not portraits. While we only display a few of these themes below, the results are

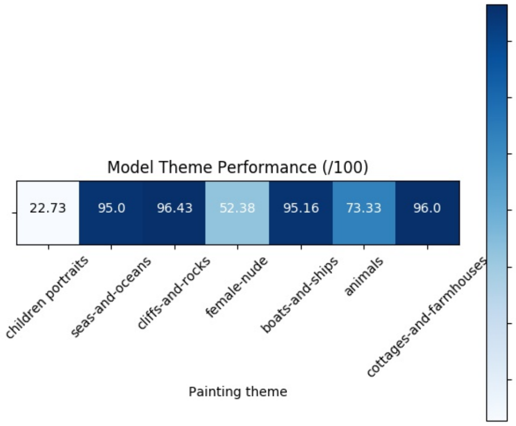

Figure 9. Correctness of our model over various other themes

telling.

Specifically, you can see that categories with little resemblance to adult portraits, such as ”seas-and-oceans”, ”cliffs-and-rocks”, and ”cottages-and-farmhouses”, all scored above 95% / 100, meaning our model correctly classified these paintings as not portraits. However, for other themes such as ”children portraits”, ”female-nude” and ”animals”, our model had more trouble. We predict that this is because these paintings either contain prominent illustrations of people in the foreground, similar to a portrait, or have other beings (e.g. animals) that simulate a portrait-like scene by having a central figure present. For children portraits in particular, it is difficult to distinguish age, especially with paintings that can be abstract.

However, this chart also underscores the subjectivity of our dataset. For instance, there are likely ”female-nude” paintings that could also classify as portraits. or ”seas-and-oceans” paintings that contain portrait-like depictions of individuals on boats. We dug into more detail about what images our model incorrectly classified and why.

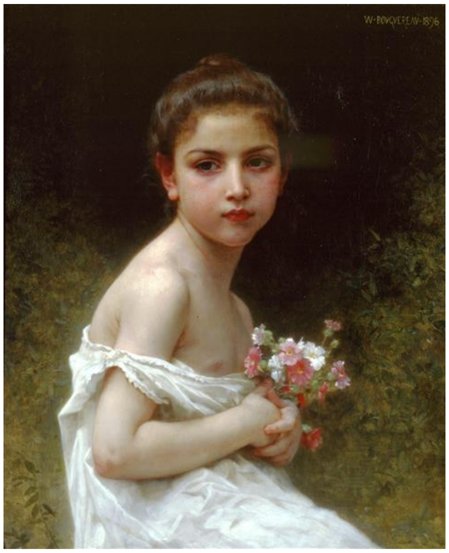

Figure 10. An image our model incorrectly thought was a portrait

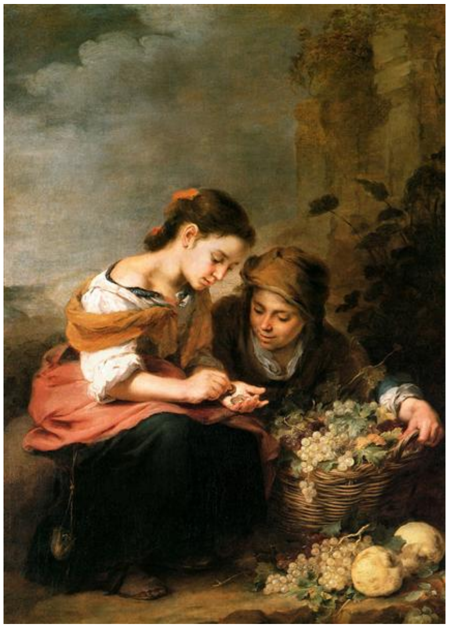

Figure 11. An image our model incorrectly thought was a portrait

5.4. Case Study 1: Misclassification as a Portrait

Figure 10 is an example of a painting that could well be considered an ”adult-portrait” if the person were slightly older, however as it stands it’s difficult to estimate the age of the subject due to the artistic and stylistic interpretations present. It underscores both the difficulty in estimating age, as well as difficulty in finding a dataset with comprehensive theme annotations, as oftentimes themes are not exclusive.

5.5. Case Study 2: Misclassification as a Portrait

Figure 11 is an example of another painting that could well be considered an ”adult-portrait” to an ordinary viewer, but the discerning eye will notice that the focus of this painting is on what is in the girl’s hand and not on the subject’s themselves. We estimate that our model was unable to pick up on this focus area, instead thinking that this painting was focusing on the individuals themselves and therefore was an adult portrait.

5.6. Case Study 3: Misclassification as not a Portrait

Figure 12 is an example of an extremely abstract ”portrait” that our model could not correctly classify. This example underscores the diversity of adult portraits present in art, and the wide range of artistic styles and interpretations present. Again, a discerning human eye can

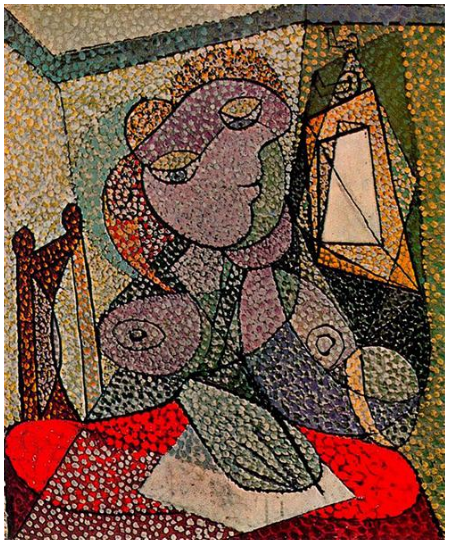

Figure 12. An image our model incorrectly thought was not a portrait

barely make out the curves and lines of a body and face centered on the canvas, in what appears to be a bedroom, but it is difficult for our model to make these distinctions due to the large artistic leaps required.

5.7. Additional Remarks on Themes

Our results here underscored the relative success our model had in discerning when people were the central subjects of paintings, but also the subjectivity of theme labeling. In many of the paintings we viewed when going over our model's classifications, we found ourselves agreeing at times with the classification as a possible ”alternate interpretation” of the theme. For this reason, we hypothesize that a more fully-labeled dataset with multiple theme labels per painting would further improve our performance.

In fact, we performed some previous work in this area in an earlier form of this project that validates this hypothesis. Specifically, we originally were attempting to build a model to classify whether a *pair of paintings* had the same theme or not. In this model, despite our extensive work with hyperparameter search, featurizers similar to the ones listed earlier, and other experiments, we failed to achieve more than 50% accuracy (close to random performance) on this dataset, even with 10K pairs of paintings. The confusion matrix in Figure 13 from that experiment underscores the subjectivity of the themes present.

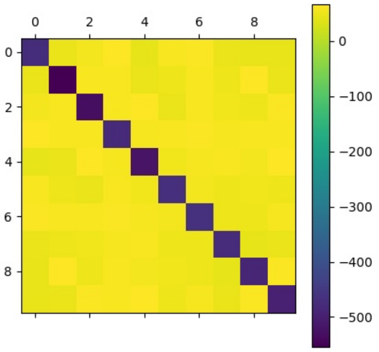

Figure 13. Scores where +1 is a correct classification of that theme pair, -1 is incorrect

Themes are indexed as the following: ['roads-and-vehicles', 'forests-and-trees', 'male-portraits', 'flowers-and-plants', 'mountains', 'boats-and-ships', 'houses-and-buildings', 'female-portraits', 'fruits-and-vegetables', 'animals'].

Overall, despite the subjectivity of high level painting themes, we were happy with our performance and the intuitive explanations present for our models' strengths and shortcomings.

6. Conclusion

In our project, we explored the possibility of the thematic classification of paintings with 4 variations of VGGNet-16: vanilla VGGNet-16, VGGNet-16 with a color histogram input featurizer, VGGNet-16 with Gram matrices for featurizing style, as well as VGGNet-16 with a fusion of color histogram and Gram matrix features. Using only two classes (adult portrait vs. non-adult-portrait), we were able to show that the color histogram and Gram matrix features were able to perform best with them together rather than separately; as a matter of fact, each feature performed a small step below the vanilla VGGNet-16 model when used separately. This is likely because color and style are both properties of a painting that are so-closely coupled that it is difficult to account for a single attribute (either color or style) without accounting for both in the model input.

As for future work, it would be interesting to implement interleaved input features throughout the layers of the network, and gather more comprehensive theme-labeled data. We would also have liked to explore more the variances within portraits across time periods. We hypothesized that there may be a way to first classify a painting according to time period, and then classify it thematically within that time period. This may help reduce stylistic variance over time and let our model learn a more concentrated definition of a theme such as ”adult-portrait”. Finally, we would have liked to have more time to explore object detection in paintings and using that as an additional featurizer. We believe that identifying objects in paintings, especially using a model trained on more abstract depictions of objects, would aid our model in identifying the focus of a particular painting and categorizing it better thematically.

Overall, we enjoyed exploring the utility of VGGNet and other image-focused machine learning algorithms on the more abstract area of paintings, particularly adult portraits, and expect much more work in this area in the future.

References

[1] The cifar-10 dataset, 2009. Accessed `https://www.cs.toronto.edu/ kriz/cifar.html/`. 3

[2] A brief report of the heuritech deep learning meetup, 2016. Accessed `https://blog.heuritech.com/2016/02/29/a-brief-report-of-the-heuritech-deep-learning-meetup-5/`. 2

[3] Wikiart.org - visual art encyclopedia, 2017. Accessed `https://wikiart.org/`. 3, 4

[4] Y. Bar, N. Levy, and L. Wolf. Classification of artistic styles using binarized features derived from a deep neural network, 2016. Supplied as additional material `http://cs231n.stanford.edu/reports/2016/pdfs/200_Report.pdf`. 1

[5] A. Bergamo, L. Torresani, and A. W. Fitzgibbon. Picodes: Learning a compact code for novel-category recognition. *Advances in Neural Information Processing Systems*, pages 2088–2096, 2011. 2

[6] J. Chen, T. N. Pappas, A. Mojsilovic, and B. E. Rogowitz. Adaptive perceptual color-texture image segmentation. *IEEE Tr. Image Proc.*, 14(10):1524–1536, 2005. 2

[7] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. *CVPR*, 2009. 2

[8] L. A. Gatys, A. S. Ecker, and M. Bethge. A neural algorithm of artistic style. *arXiv:1508.06576*, 2015. 2

[9] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, pages 1097–1105, 2012. 1, 2

[10] W. Y. Ma, Y. Deng, and B. S. Manjunath. Tools for texture/color based search of images. *Human Vision and Electronic Imaging II*, 3016:496–507, 1997. 2

[11] A. Mojsilovic, J. Kova, J. Hu, R. J. Safranek, and S. K. Ganapathy. Matching and retrieval based on the vocabulary and grammar of color patterns. *IEEE Tr. Image Proc.*, 1(1):38–54, 2000. 2

[12] J. Parker. Algorithms for image processing and computer vision. *John Wiley Sons*, 2010. 1

[13] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *ICLR*, 2014. 1, 2

[14] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. *NIPS*, 2014. 1, 2

[15] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. *Computer Vision–ECCV 2010*, pages 776–789, 2010. 2

[16] M. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. *ECCV*, 2014. 1, 2

[17] J. Zujovic, L. Gandy, S. Friedman, B. Pardo, and T. Pappas. Classifying paintings by artistic genre: An analysis of features classifiers. *Multimedia Signal Processing*, 2009. 1, 2