Binary Classification of Beautiful vs. Non-Beautiful Photographs Using A Deep Convolutional Neural Network

Stephen Crawford

Gibbons Research Fellowship, Bowdoin College, Brunswick, ME.
Email: scrawfor@bowdoin.edu

Aesthetics describes the field of study related to the appreciation or definition of beauty. The discussion of aesthetics generally resides in the fields of philosophy and art [1]. However, this work explores a mathematical explanation of aesthetics through the creation of a binary classification model tasked with differentiating between beautiful photographs and non-beautiful photographs. In order to train the model, a database of award-winning and non-award-winning photos was collected. Award-winning photos represented the beautiful while non-award-winning photos represented the non-beautiful. A densely connected model was trained on features extracted by running the data through a VGG16-based feature extraction function. Results suggest that it is possible to differentiate between the beautiful and non-beautiful but that more specific datasets are easier to differentiate within.

Keywords: Deep Learning; Convolutional Neural Network; Confusion Matrix; Binary Classification; Feature Extraction

1. INTRODUCTION

This work describes the findings of research conducted with the aim of creating a binary classification model for differentiating beautiful vs. non-beautiful photographs. The purpose of this research was to use a deep convolutional neural network (DCNN) to create a classifier which would then be analysed for insight into the differences between beautiful photographs and non-beautiful photographs. This insight is important for the answering of the philosophical question "what is beauty?", the valuable guidance it offers in how artists may create beautiful works, and as further research into a popular line of inquiry.

When creating models to analyze photo quality, there are two main lines of inquiry. There are works which are primarily focused on the correction of images with respect to some distortion [2, 3, 4, 5], and works which are interested in evaluating how "good" photos are [6, 7, 8]. This work falls in the latter category.

The numerous works focusing on the correction of distorted images generally make use of reference images or knowledge of the types of distortion present [9, 10]. There are also works such as Kang *et al.* which have dealt with the more challenging area of No-Reference Image Quality Assessment (NR-IQA) [2, 3, 5, 11]. This should be differentiated from the purpose of this and other research which is not focusing on Image Quality Assessment but instead focuses on what will hereto be referred to as Photo Quality Assessment (PQA). The distinction being that the latter group of works focus on evaluating how "good" a given image is in general, not on how to improve the quality of a specific image or on determining metrics such as sharpness or clarity.

PQA works have often made use of photographic composition rules such as the rule of thirds, contrast, and exposure levels [6, 12, 13]. The work discussed in this paper instead takes a more laissez-faire approach to how the model interprets the features of the photographs. Given it has already been shown that a model can be trained to evaluate a photograph when high-level features are designed, this work uses what can be thought of as base features (the photo themselves) in an attempt to improve on the results of previous low-level work [6, 14] while avoiding the overhead associated with the high-level approaches. To clarify, low-level features are qualities of images such as contrast and exposure whereas, high-level features are generally features designed by researchers for the specific task of classifying the images. High-level features usually consist of multiple low-level features combined. An example of a high-level feature is color distribution [12]. Research by Dong *et al.* is likely the most similar work. Dong also ran photos directly through a convolutional neural network trained on ImageNet. However, that work relied on an older model architecture and use of a "spatial pyramid" to rectify image size differences [8].

Another major difference between this work and

associated research is the nature of the dataset used. Other works have generally dealt with publicly sourced datasets such as *Photo.net* [6] or a DPChallenge.com dataset collected by Ke *et al.* [12, 13]. For both databases mentioned, the photos were submitted by public users. *Photo.net* is a community site where users can upload and rate the photos. DPChallenge.com is a photography community forum and challenge site where users can participate in weekly challenges aimed at improving their skills.

Dong took a more thorough approach in using images from the CUHKPQ dataset and the Aesthetic Visual Analysis dataset [8]. Though Dong's databases are arguably better representations than the *Photo.net* dataset or the database collected by Ke, they still involve classification by volunteer reviewers. Although there are certainly beautiful, professional-level photos in all the databases, the photos used in this research arguably represent a more accurate ground-truth for beautiful photos. This will be discussed further in the following section.

2. METHODS

2.1. Data Collection

In order to train a model to differentiate between beautiful and non-beautiful photographs, it was necessary to have ample samples of both classes. As mentioned, previous research has approached the issue of data collection in a number of ways. The most relevant work to this research [6, 8, 12, 13] saw datasets aggregated from user submitted photo sites. In the datasets used, the researchers made the assumption that there was a certain disparity of quality in the photos on the site.

This assumption may not be true however. While there are both professionals and amateurs on the two sites, both sites are tailored towards photography enthusiasts. The sources were two websites where individuals are more likely to be posting "good" photos regardless if they have made a career out of photography or not. There are certainly better and worse photos amongst the two datasets, but it does not mean that there will be enough truly beautiful and non-beautiful photos for the aesthetics of the photos to be learned by a model. Instead the model may simply differentiate amongst a database of mostly above-average photos. Because aesthetics is the topic at hand, only true beauty as recognized by expects should serve as the ground truth labels. The peer-review sourced datasets lack this confidence.

Likewise, Datta *et al.'s* mentions that users are likely rating numerous photos in a short period and that, peer-reviewed data generally has a significant amount of noise. This being the case, it seems unlikely the work would generalize to a higher standard of judging such as an art critique or awards panel. While there are highly and poorly rated photos on Photo.net, the nature of the site makes it unlikely that these photos were judged objectively and without explicit or implicit comparison to other photos on the site. As such, while the low ranking photos are likely worse than the higher rated ones, they are not necessarily non-beautiful.

In order to create a serviceable proxy for the two classes of beautiful photos and non-beautiful photos, an alternative approach was adopted. If photos were arbitrarily selected as beautiful or not by the investigator, then there would be a great deal of bias to the dataset. Furthermore, the common practice of data augmentation could not be used. Though it is often applied to small datasets, because data augmentation executes transformations on the samples, it may have caused images to transform across classes. During data augmentation, a beautiful photo may become non-beautiful.

In order to create as representative dataset as possible, photos which had won recognition in any of six prestigious photography contests were used as stand-ins for beautiful photos. Justifiably, it was presumed that a photo which had won recognition in a major contest would have been deemed by experts to be beautiful. The six contests to be used are the Deutsche Börse Photography Foundation Prize; the World Press Photo Contest; the Hasselblad Award; the Sony World Photography Awards; the National Geographic Photo Contest; and the Fine Art Photography Awards.

From these contests, a dataset of 12,000 award-winning photos was created representing beautiful photos. For each of these contests, photos were downloaded in their original format. For contests dating back to black and white photography, only the years which were in majority color were used. This decision was made in order to prevent any artificial bias on color characteristics based on the general adoption of color photography around 1980.

In order to build the non-award-winning portion of the dataset, a simpler approach was employed. Again to avoid bias incurred by manually selecting non-beautiful photos, photographs were randomly extracted from an existing database. Specifically, the first 200,000 photos of the LabelMe database were downloaded before a random subsample of 12,000 images was selected. This was a quick and efficient method of building the non-award-winning portion of the database.

The collection of the non-award-winning photos from the LabelMe database was determined to be the best option for acquiring a proxy of non-beautiful photos both for its convenience and the fact that many of the photos are of everyday objects which seem unlikely to be the subject of an award-winning photo. While there are some exceptions, many photos in the LabelMe database are of mundane scenes.

Finally, in order to create dataset splits for training, validation, and testing, an equal number of photos from each of the two classes were randomly assigned to each split. Specifically, 6000 award-winning photographs and

Sample Type	Training	Validation	Testing
Non-Winning	6000	3000	3000
Winning	6000	3000	3000

TABLE 1. The number of samples for each split.

6000 non-award-winning photographs were assigned to the training split. 3000 photos of each class were then assigned to the validation and test splits (*Table 1*).

2.2. Model Architecture

After collecting a suitable dataset, the model architecture could be selected. A fundamental principle behind deep learning is the use of calculus to train weights via back propagation [15]. One specific type of deep learning model is the convolutional neural network (CNN). CNNs provide state-of-the-art performance in object recognition tasks and have a wide variety of applications in the computer vision and robotic sensing subfields [16, 17, 18].

The effectiveness of CNNs in classifying images is in large part due to their ability to learn localized patterns within data and then generalize these patterns across dimensional translations [19]. When CNNs train on data, the layers they consist of learn to recognize progressively higher level representations. For instance, a first layer in a CNN may learn to identify horizontal lines while a later layer may search images for triangles (*Figure 1*).

A result of CNNs' ability to generalize localized patterns within data is the ability to use feature extraction. Feature extraction is a method of model training whereby the convolutional base of a model trained on separate data is used to identify components of the target samples [19]. The use of feature extraction is common when there is relatively little data.

Because CNNs are able to generalize local patterns within images as well as make use of layers trained with other datasets, any notable trends in the award-winning vs. non-award-winning photographs should be detected [19].

Given the success of feature extraction when operating with smaller datasets, it made the most sense to use the convolutional base of a preexisting network model. In this case, the VGG16 model trained on the ImageNet database was a good candidate. VGG16 uses small convolutional filters and easily outperformed its competitors when it was developed. Furthermore, the architecture was found to generalize well to other datasets [20]. The model, has a total depth of 23 layers and 138,357,544 total parameters. [21] The model used the weights learned during training on an ImageNet subset dataset of 1,430,000 images as opposed to the total ImageNet dataset of 14,200,000 images [20, 22].

Importantly, the top layers of the VGG16 convolutional base were disabled so that the fully connected layers of the base did not function. Using only the convolutional base, feature extraction was possible by running the samples of shape `(150, 150, 3)` and their labels through the convolutional layers and then storing the output in a tensor of shape `(batch_size, 4, 4, 512)`. This output tensor is then fed into a secondary densely connected classifier which differentiates the extracted features between the two target classes. The densely connected classifier consists of a sequential model with 2 layers. The first layer cxontains 512 hidden units and makes use of a Rectified Linear Unit (ReLU) activation function. It has a 0.5 dropout rate to mitigate overfitting. The final layer is a mono-nodal classifier with a sigmoid activation. The model uses an Adam optimizer with learning rate=4e-4 and a binary crossentropy loss function.

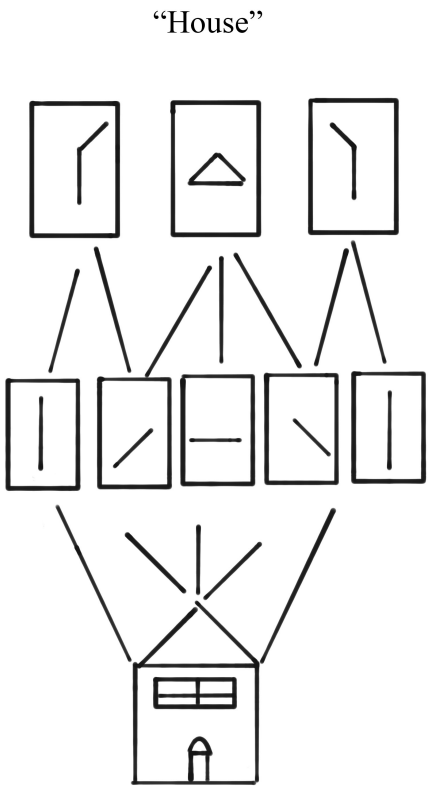

FIGURE 1. An example of how convolutional layers breakdown image features.

A classification approach was chosen because previous work found that classification models better evaluate photo quality than regression models [6].

3. RESULTS

For this research, the model was to be considered effective if it could differentiate between winning and non-winning photos at a rate higher than 0.5 or 50%. Given the even dataset, the model could be expected to guess the label of a sample correctly 50% of the time

without knowing anything about the sample.

With respect to this naive baseline, all version of the model performed well. Even in early iterations of the model where there was significant overfitting, the model still maintained an accuracy above 50%.

Results for the model were heavily dependent on the exact configuration of the densely connected classifier and the training parameters. Because the dataset was relatively small and data augmentation could not be relied upon, overfitting was a significant issue when the densely connected layer had too many hidden units. In order to verify the functionality of the model, a confusion matrix was manually calculated. For each iteration of the model, the precision, recall, and F-measure were calculated by tallying the results of the model's predictions over the 6000 sample test set (*Figure 2*). The confusion matrices show that the model performed better than the baseline. It performed worse than the models which used high-level feature extraction to train their models [7, 12, 13].

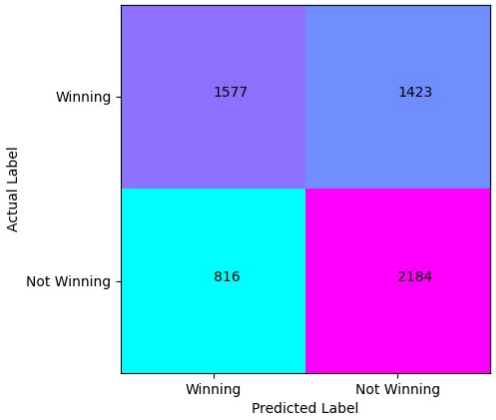

FIGURE 2. A confusion matrix representing the results of the model trained on all photos.

Interestingly, it appears that the model struggles to classify the same images that a person may have difficulty classifying. In a given sample, the classifications it mislabels are often those which are somewhat ambiguous or won awards from the Fine Art Photography Contest where more abstract submissions are allowed (*Figure 3*).

It is important to note that the built-in `model.evaluate()` function does not correctly represent the results of the model after training. Presumably because of the feature extraction, the model is said to have effectively no loss and near perfect accuracy. This was initially thought to be a result of overfitting, but the modification of the model to include further overfitting prevention failed to deliver noticeable changes in the results of `model.evaluate()`. Upon failing to correct this abnormality, the `sklearn.metrics.classification_report()` function was used instead. This function calculated all reported statistics which were also manually confirmed (*Table 2*).

	f1-score	precision	recall	support
Non-Winning	0.66	0.61	0.73	3000.00
Winning	0.58	0.66	0.53	3000.00
Macro Average	0.62	0.63	0.63	6000.00
Weighted Average	0.62	0.63	0.63	6000.00
Accuracy	0.63			

TABLE 2. Results from training on all photos.

	f1-score	precision	recall	support
Non-Winning	0.83	0.75	0.94	148.00
Winning	0.79	0.92	0.69	152.00
Macro Average	0.81	0.83	0.81	300.00
Weighted Average	0.81	0.84	0.81	300.00
Accuracy	0.81			

TABLE 3. Results from training on architecture photos.

In order to further validate the model, the same architecture was trained on an entirely new dataset. While the main dataset included photos across many categories, the second dataset included only architecture photos. This smaller dataset included 900 award-winning photos and 900 non-award-winning photos. The winning photos were taken from the Sony World Photography Awards while the non-winning photos were from a dataset created by Xu *et al.* [23]. The division of the smaller dataset was 600 photos of each class for training and 150 photos of each class for both validation and testing. The process of training and analyzing the model was the same as before.

Significantly, the model trained on the smaller but more specific dataset out performed the model trained on the more diverse dataset. All statistics of the model trained on only architecture photos show better performance than the model trained on the entire set of award-winning photographs (*Table 3* & *Figure 4*).

4. DISCUSSION

The models struggled to classify the photographs when they were taken from a wider range of samples. It was easier to train the same architecture to differentiate between the beautiful and non-beautiful architecture photos than it was to train the model to differentiate between all beautiful and all non-beautiful photos. This is to be expected considering what is known about how convolutional networks work. Plainly, when there are mostly buildings in the photographs, it should be easier for the model to identify patterns of features.

Likewise, the dataset with only architecture photos removes many of the photos which appear to have caused the most problems for the original model. The award-winning photographs from the Fine Arts

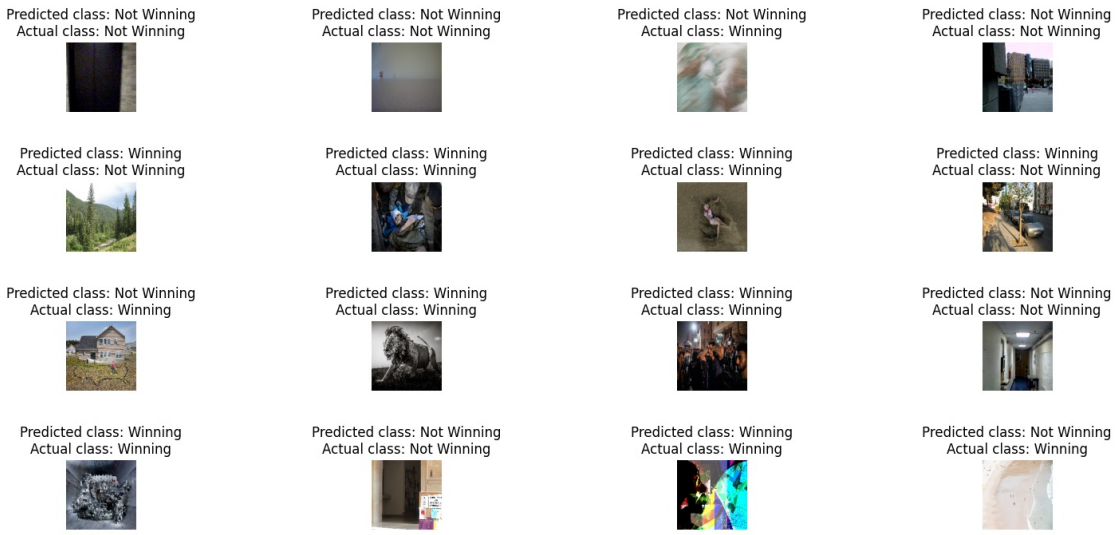

FIGURE 3. A sample of the predictions made by the model for a set of photographs.

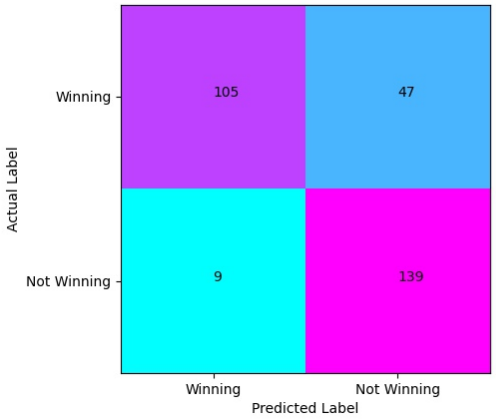

FIGURE 4. A confusion matrix representing the results of the model trained on just architecture photos.

Photography Contest were particularly prone to being mislabeled by the classifier. The fact that these photos were abstract and rarely contained identifiable objects would have made the convolutional base of the model struggle to recognize the types of patterns it learned when analyzing everyday objects via ImageNet. This suggests that there is either a more obvious difference between an award-winning architecture photo and a non-award-winning architecture photo or more likely, classification becomes much more difficult when there are many different types of photos.

These findings were expected considering the architecture of the model and the scarcity of data. While feature extraction is logical given the dataset consisted of only 24,000 images, using a convolutional base which was previously trained on identifying everyday objects likely led to some of the difficulties the model had generalizing the abstract photographs. If the model instead had a convolutional base which was trained solely on the problem at hand, it may have more success recognizing an intentionally abstract photo when compared to a photo which is simply out of focus [8].

The use of the convolutional base from the VGG16 model was necessary in this case because of the scarcity of available data. Before this research, it was believed there was no public database containing so many award-winning photographs. While there are the databases used in the aforementioned works, as stated, it is believed that a more representative database was necessary.

When an original model was created without any feature extraction, the predictions offered by the model were in-line with a random guess and the loss of the model remained very high. The model was unable to successfully learn because there were so many different features to be learned on relatively little data. Even with the 12,000 award-winning photos, a useful classifer could not be trained. As has been shown by other works, far superior classification rates are possible through more advanced high-level feature extraction [6, 12, 13].

5. CONCLUSIONS

Further research should focus on improving the accuracy of the model as well as identifying the key differences between award-winning and non-award-winning photos. While CNNs and neural networks in general are often described as "black boxes," it is possible to visualize the activations of CNN layers. Using techniques such as gradient class activation mapping (GRAD-CAM), it is possible to locate the

portions of images which most strongly cause the model to label a given image as a specific class. A common method of using GRAD-CAM is to overlay a grid of where the model is predicting a specific class over the top of the original image. The overlay creates a heatmap where it is possible to see the parts of the image which the model most strongly identifies as the target class. In the case of this research, this would allow future investigators to determine which parts of an image made it beautiful or non-beautiful.

Using visualizing techniques, there is a great deal to be discovered about the mathematical basis of beauty and aesthetics. Past work has shown that it is possible to achieve very accurate models when using high-level features to train [6, 12, 13]. This work has shown that base features underperform compared to their high-level counterparts, however the improvement in accuracy shown when training on just the architecture dataset suggests that with more specific data, better results may be possible.

ACKNOWLEDGEMENTS

This research was conducted under funding provided by the Gibbons Research Fellowship and its sponsors. The investigator appreciates those who made this work possible and specifically would like to thank Professor Sean Barker for offering advice.

REFERENCES

[1] *The Camdridge Dictionary*. The Cambridge Press.

[2] Zhang, W., Ma, K., Yan, J., Deng, D., and Wang, Z. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, **30**, 3647.

[3] Bosse, S., Maniry, D., Wiegand, T., and Samek, W. A deep neural network for image quality assessment. *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 3773–3777.

[4] Bosse, S., Maniry, D., Mller, K.-R., Wiegand, T., and Samek, W. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, **27**, 206–219.

[5] Ma, K., Liu, W., Zhang, K., Duanmu, Z., Wang, Z., and Zuo, W. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing*, **27**, 1202–1213.

[6] Datta, R., Joshi, D., Li, J., and Wang, J. Z. Studying aesthetics in photographic images using a computational approach. In Leonardis, A., Bischof, H., and Pinz, A. (eds.), *Computer Vision – ECCV 2006*, Berlin, Heidelberg, pp. 288–301. Springer Berlin Heidelberg.

[7] Li, C., Loui, A., and Chen, T. Towards aesthetics: A photo quality assessment and photo selection system, . 10, pp. 827–830.

[8] Dong, Z., Shen, X., Li, H., and Tian, X. Photo quality assessment with dcnn that understands image well. In He, X., Luo, S., Tao, D., Xu, C., Yang, J., and Hasan, M. A. (eds.), *MultiMedia Modeling*, Cham, pp. 524–535. Springer International Publishing.

[9] Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, **13**, 600–612.

[10] Sheikh, H., Bovik, A., and de Veciana, G. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on Image Processing*, **14**, 2117–2128.

[11] Kang, L., Ye, P., Li, Y., and Doermann, D. Convolutional neural networks for no-reference image quality assessment. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.

[12] Ke, Y., Tang, X., and Jing, F. The design of high-level features for photo quality assessment, . 07, pp. 419–426.

[13] Luo, Y. and Tang, X. Photo and video quality evaluation: Focusing on the subject, . 10, pp. 386–399.

[14] Tong, H., Li, M., Zhang, H.-J., He, J., and Zhang, C. Classification of digital photos taken by photographers or home users. In Aizawa, K., Nakamura, Y., and Satoh, S. (eds.), *Advances in Multimedia Information Processing - PCM 2004*, Berlin, Heidelberg, pp. 198–205. Springer Berlin Heidelberg.

[15] LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature*, **521**, 436–444.

[16] Sahoo A.K., D. H., Pradhan C. *Nature Inspried Computing for Data Science*. Springer, Cham.

[17] Abbas, A., Abdelsamea, M., and Gaber, M. Classification of covid-19 in chest x-ray images using detrac deep convolutional neural network. *Appl Intell*, **51**.

[18] Gu, Y., Li, Z., Zhang, Z., Li, J., and Chen, L. Path tracking control of field information-collecting robot based on improved convolutional neural network algorithm. *Sensors*, **20**.

[19] Chollet, F. *Deep Learning with Python*, 1st edition. Manning Publications Co., USA.

[20] Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition.

[21] Chollet, F. Vgg16 and vgg19.

[22] Fei-Fei, L., Deng, J., Russakovsky, O., Berg, A., and Li, K. Imagenet.

[23] Xu, Z., Tao, D., Zhang, Y., Wu, J., and Tsoi, A. Architectural style classification using multinomial latent logistic regression. *ECCV*.