Modeling Image-to-Image Confusions in Memory

by

Anthony Dong Zhao

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2015

Author . . . .
Department of Electrical Engineering and Computer Science
May 22, 2015

Certified by . . . .
Aude Oliva
Principal Research Scientist
Thesis Supervisor

Accepted by . . . .
Prof. Albert R. Meyer
Chairman, Masters of Engineering Thesis Committee

Modeling Image-to-Image Confusions in Memory

by

Anthony Dong Zhao

Submitted to the Department of Electrical Engineering and Computer Science
on May 22, 2015, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Computer Science and Engineering

Abstract

Previous experiments have examined what causes images to be remembered or forgotten. In these experiments, participants sometimes create false positives when identifying images they have seen before, but the precise cause of these false positives has remained unclear. We examine confusions between individual images as a possible cause of these false positives. We first introduce a new experimental task for examining measuring the rates at which participants confuse one image for another and show that the images prone to false positives are also ones that people tend to confuse. Second, we show that there is a correlation between how often people confuse pairs of images and how similar they find those pairs. Finally, we train a siamese neural network to predict confusions between pairs of images. By studying the mechanisms behind the failures of memory, we hope to increase our understanding of memory as a whole and move closer to a computational model of memory.

Thesis Supervisor: Aude Oliva
Title: Principal Research Scientist

Acknowledgments

Zoya and Phillip, for working closely with me for the past two years as they helped me turn what started as a UROP project into this thesis. I learned an incredible amount while working with them and, until recently, never thought I'd be making a real, significant contribution to the body of research.

Tim Brady, who gave us advice and information that helped with our research and writing.

My advisor Karren Sollins, for offering advice and encouragement throughout my time at MIT.

Everyone in Next House 2 West, and all its frequent visitors, for being my community for four years and never giving me a dull moment.

My parents, for putting up with me for 20+ years and paying for my tuition. I don't intend to let their support and faith in me go to waste.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Existing research on memory [13, 3, 2] has shown that people consistently remember the same images and forget the same images. This means that memorability is a property intrinsic to the image, and that memory is a predictable process that can be potentially be modeled. However, if memory can be modeled, that would mean its failures can be predicted and modeled as well.

In this paper, we (1) introduce a new experimental task designed to determine which images are most frequently confused with each other, (2) report that participants tend to confuse the same pairs of images that they find visually similar, (3) report that image-to-image confusions explain a good amount of the variance in false alarm rates, (4) show that siamese neural nets can be used to predict how likely two images are to be confused, and (5) show that they can also be used as a tool allowing us to compare confusions between images with semantic similarity.

Failures of memory can be divided into false positives and false negatives. A false negative occurs when a participant simply forgets that he or she has seen an image, while false positives occur when a participant falsely believes he or she has previously seen a given image. Existing research on photo memorability has largely ignored false positives, choosing instead to focus on what causes images to be remembered or forgotten [13, 3, 2, 12, 6]. This paper is an exploration of false positives, intended to fill this gap in the literature.

We gathered data on confusions and similarity of pairs of images drawn from the

SUN dataset [28]. We show correlations of the confusion and similarity data with each other and with existing measures of image memorability. We also construct predictive models for confusions and similarity, and discuss what these models tell us about how and why images are confused.

1.1 Motivation

Having a good model of confusions and failures of memory will provide an understanding of the memory mistakes people make. More generally, it is also an exploration into how the human brain encodes images. Knowing what kinds of images are likely to be confused is useful when teaching concepts or presenting images to people, as it provides information on what images can be shown together, and which should not be due to high likelihood of confusions. This is similar to the motivation of [3].

Knowing the causes of confusions may also shed some light on the phenomena of déjà-vu and false memories. This area of research is beyond the scope of this paper, but is relevant to applications such as the evaluation of eyewitness testimony.

Chapter 2

Background and Related Work

2.1 Image Memorability Experiments

In a 2011 paper, Isola et al. tested the memorability of images using a sequential memory game [13]. The game presented participants with a sequence of images, shown one at a time for 1s with a short delay between images, and instructed them to provide a response if the image was one that had appeared previously. For a repeated image, if the participant correctly indicated that the image has appeared before, that event is called a *hit*. Otherwise, if the participant failed to provide the response, that event is called a *miss*.

The paper calculated a memorability score for images based on the ratio of hits to total repeat appearances of that image (across participants), or *hit rate* (**HR**), and found that memorability had high consistency across participants. When participants were split into two groups at random, the memorability scores of the two groups were closely correlated, meaning that the scores of one group were a good predictor for the scores of the other. This suggests that people tend to remember and forget the same images.

An additional memorability metric explored was based on *false alarms*, which are events where the participant indicated that an image had previously appeared when that image appeared only for the first time. The *false alarm rate* (**FAR**), or the ratio of false alarms to total appearances, also had high cross-participant consistency, meaning that people are similar in terms of what images they incorrectly believe they have seen.

Later studies [3, 2] used Isola's sequential memory game setup with different image sets ([3] used graphs, and [2] used faces) and also found high cross-participant consistency for both HR and FAR, indicating a generalization of the findings across different visual categories.

These results form a body of evidence showing that people tend to make similar mistakes, indicating that memory is a predictable process that can potentially be modeled. However, while the existing research shows what mistakes people make, it does not show why people make those mistakes, in terms of which pairs of images are being confused.

2.2 Factors affecting Memorability

The memorability of an image is influenced by both intrinsic and extrinsic factors, where intrinsic factors are those based solely on the contents of the image, and all other factors are considered extrinsic.

2.2.1 Intrinsic Factors

Consistency in memorability scores indicates that some images are inherently more memorable than others, as determined by their contents [13, 12]. Simple low-level perceptual features like hue, saturation, and value of pixels are poor predictors of memorability, while semantic features based on the presence of objects in the image are more predictive. Some objects, such as people and cars, were correlated with higher memorability, while other objects, like mountains and trees, were correlated with lower memorability [13].

Recently, Kholsa et al. showed that a deep neural net can do very well at predicting an image's memorability directly from its pixels [15]. This was done by adapting a neural net originally trained to label objects in images. The success of this method gives further evidence that the object-level content of an image is predictive of whether or not it will be remembered.

2.2.2 Extrinsic Factors

An image's memorability is affected by its context, defined by the images that appear together with it (in a sequence) [6]. Prior work has shown that human memory can accurately store information about a large number of images if they are sufficiently distinct from each other [4], and that performance becomes worse when images are instead drawn from a pool of similar objects [17].

Furthermore, the most memorable images in one context are not necessarily the same as the most memorable images in a different context [6]. Some images that are memorable within their own scene category, for instance, have near-neighbor images in other categories, and have lower mixed-category memorability as a result. Also, the more similar images there are in a given context, the fewer images will be remembered overall. People can remember more images in a single sitting when all those images are more distinct from each other.

2.3 The Similarity Space of Memory

The literature [13, 3, 6, 2] focuses on the memorability of single images, examining what characteristics of an image make it more likely to be remembered when it is seen again. However, these experiments have largely ignored false alarms. Recall that a false alarm occurs when the participant sees an image and incorrectly believes that it has occurred before. This implies that the participant is confusing that image for some previous image (or images), but the setup of the sequential memory game makes it impossible to determine which of the previous images in the sequence are serving as the source of confusion. Isola et al. found that simple image features, like the average hue of an image, and non-semantic features, like the number of objects shown, were poor predictors of memorability [13]. The work of Brady et al. [5] offers further support by showing that the color of an object is forgotten more quickly than the state of the object (e.g. a cabinet being open or closed). Thus, semantic features are more strongly represented in memory than perceptual ones. It may be that images are converted to a representation in "memory space," and that failures of memory occur when two images have similar representations in memory, resulting in collisions and interference. This model is supported by a large body of prior work [6, 17, 3, 11, 23] showing that people make more mistakes when the images or other stimuli are more similar to each other and that they are better at remembering and recalling stimuli that stand out from their context. This literature suggests that failures of memory, and therefore confusions between two images, are also predictable and modelable processes. This paper seeks to explore and predict these confusions, and to find out what characteristics - perceptual or conceptual - will cause two images to be confused with each other, leading to false alarms.

2.4 Convolutional Neural Nets for Visual Tasks

Recent work [18, 19, 21] has established convolutional neural nets (CNNs) as powerful tools for image processing tasks, and we will be using them to model failures of memory. Generic neural nets must maintain a large number of neurons because their input is of indeterminate format, making them computationally expensive. In contrast, CNNs assume their input is an image, so they can be implemented to scale better and be more computationally efficient [1].

Each layer of a CNN applies a transformation to its input, such as convolution, normalization, or downsampling. Some of these layers have weights and biases, which are adjusted as the neural net is trained. Generally, a CNN uses one or more convolutional layers to extract features from the image with a series of convolutions, and then uses fully-connected layers to calculate a label or classification from those features.

Training CNNs on handwritten digit recognition [19] and image classification problems [18], has created classifiers that are competitive with human performance on those tasks. Even when trained on tasks other than the ones they were originally designed to solve, their performance is competitive with that of other methods, and this improves further when the network is tuned and specialized for a task [21]. Since these CNNs are trained with object labels, their layers are trained to extract and encode semantic content, and they learn to recognize the presence of different objects in an image. Although the task in this paper is a regression rather than classification task, the literature indicates that a CNN is likely to provide good results.

Chapter 3

Experimental Design

3.1 Stimuli

The images used for the experiments in this paper are a subset of those used in [6]. Those images were originally scenes from MIT's SUN database [28], each individually resized so its larger dimension was 256 pixels, then cut down to the central 256×256 pixels from the center of the resized image.

We used images drawn from the *airport terminal*, *amusement park*, *badlands*, and *cockpit* categories. This results in one category with high memorability (*amusement park*), one with low memorability (*cockpit*), and two with average memorability (*airport terminal* and *badlands*). This also gives us two indoor and two outdoor categories. We only sampled images for which memorability scores (HR and FAR) were measured.

3.2 Memory Card Game

In the memory card game task, the participant is initially shown a grid of hidden images represented as face-down cards. Clicking on a card turns it face-up, revealing the image assigned to that space. The participant then clicks on a second card. If the revealed images match, they are removed from the grid and replaced with check marks. Otherwise, both are turned face-down. The participant repeats this process

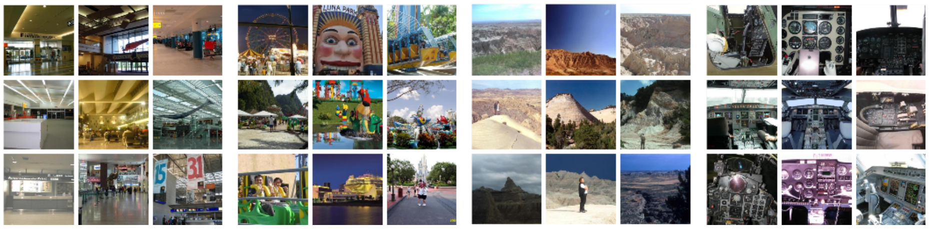


Figure 3-1: From left to right: 9 images each from the *airport terminal, amusement park, badlands,* and *cockpit* categories.

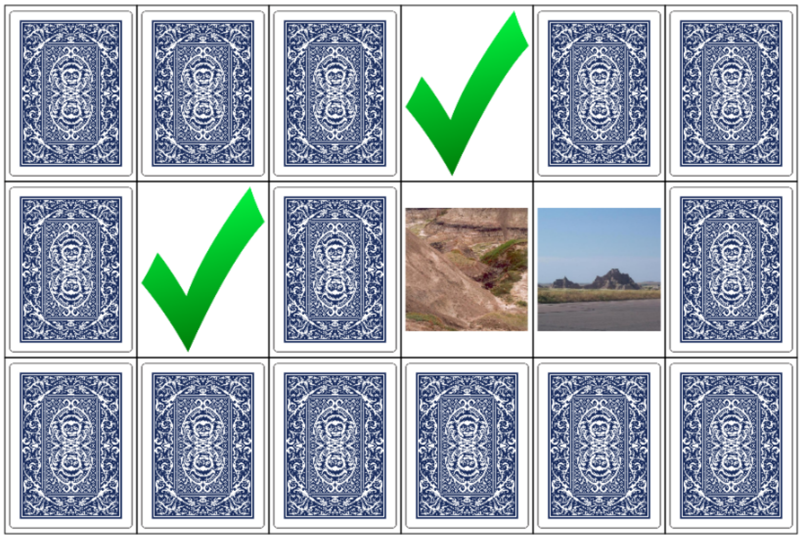


Figure 3-2: A screenshot of the memory card game in progress. Up to two images are turned face-up at a time. The check marks replace images that have been matched and removed from the grid.

until all images have been matched and removed from the grid. Each image matches exactly one other image somewhere in the grid.

We record every click that changes the game state. That is, we record any click that causes an image to be turned over, but ignore clicks outside the game grid, clicks on images that are already face-up, and clicks on images that have already been matched and removed.

There is some precedent for experimental tasks similar to this one. Kibbe, in [16], used a task where participants were presented with hidden images and instructed to find sets of images that matched or differed in certain characteristics. However, Kibbe's task was focused on the use of exploration to refresh memory, rather than confusions between images.

We designed our task so that the actions of the participants directly show us the images that they are confusing, making it easy to know the exact images that are being confused with each other. Additionally, the task's similarity to commonly played memory games means it is intuitive to participants and can be presented with a minimum of instructions.

The grid used in this task was 6 images wide by 3 images high. Each image was 140×140px and centered in a cell that was 150px wide by 200px tall. Each participant played one game. Table 3.1 shows the number of games played for each category.

Category	Games played
airport terminal	1,621
amusement park	1,621
badlands	1,615
cockpit	1,598

Table 3.1: Table showing the number of memory games played for each category.

3.3 Similarity Judgment Task

In the similarityjudgment task, the participant is shown two randomly selected pairs of images. He or she is instructed to select the pair with the images that are more similar, and the choice is recorded. Each participant repeats this process a fixed number of times, with two new pairs generated each time. The images within each pair cannot be identical, and the two pairs do not show the same two images, though the same image may appear in both pairs.

Previous work [20] has also used participants to judge similarity of images because direct comparison of perceptual features does not replicate how visual encoding simplifies parts of images, causing them to appear similar even if direct comparison would identify them as different. Previous experiments have used comparison-based tasks asking which of two test images is more similar to a reference image [20, 24], using a forced-choice experiment for its ability to estimate difference thresholds [25]. Our task uses two independent pairs of images because we wish to rank all pairs of

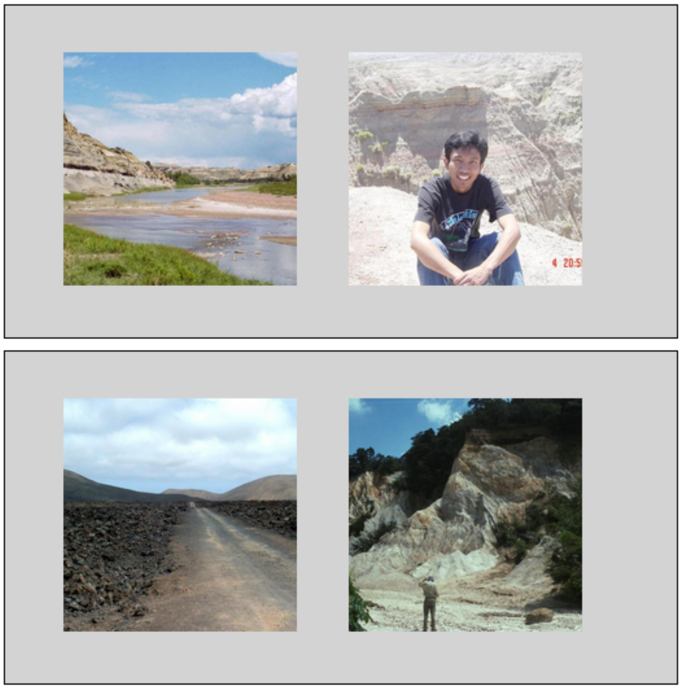


Figure 3-3: An example of two pairs presented for the subjective judgment task. Participants are instructed to choose the pair containing the images that are more similar to each other. The gray rectangles signal the participant that the pairs are grouped horizontally and not vertically.

images rather than only comparing pairs that share an image in common.

Images were shown at 200×200px. Most participants judged 25 pairs of images, but some of the earlier participants judged a different number of images as we changed this parameter. Table 3.2 shows the number of participants and number of judgments made between pairs for each category.

Category	Participants	Games played
airport terminal	850	25,655
amusement park	895	26,945
badlands	899	27,130
cockpit	899	27,085

Table 3.2: Table showing the number of participants and judgments for each category in the similarity judgment task.

3.4 Amazon Mechanical Turk Protocol

Both experiment types were posted on Amazon Mechanical Turk (AMT). Each task presents the worker with images taken from a single category. AMT workers were only allowed to participate once in a task for a given experiment type and image category, but were allowed to participate in any number of tasks. Participants were compensated for each completed task. AMT anonymizes all participant information, including names.

IRB COUHES approval was obtained for all AMT experiments. A COUHES text appears with each posted experiment on AMT, to indicate anonymity of data and permission to terminate the experiment or contact the experimenters at any time.

Chapter 4

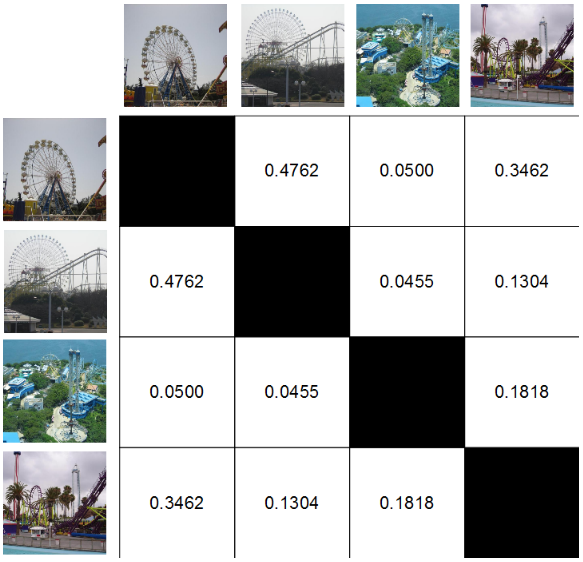
Analysis Methods and Computational Models

4.1 Generating Confusion Rates

The *confusion rate* (**CR**) for an image pair is calculated based on the clicks made by the participants in the memory game. It is assumed that, for a pair of clicks, the participant makes the first click, examines the image, and then clicks on the location of the image that he believes to match the first image. If the second image is incorrect, that means the participant has confused the second image with the first.

However, not all click pairs are relevant for our task of studying confusions between images. For any given pair of clicks, the pair represents a confusion only if the participant had previously seen the second image to be clicked. Otherwise, the participant has no memory of the second image, and cannot have confused it for anything. This means we ignore click pairs if the second click is on an image that has not previously been revealed, and we ignore only those click pairs. It is irrelevant whether or not the first click is on a previously revealed image.

The empirical CR for a given pair of images is calculated as the number of games where those two images were confused together at least once (in a click pair that is not being ignored), divided by the number of games where those two images appeared together. We also measure the *overall confusion rate* of individual images. The overall



	■	0.4762	0.0500	0.3462
	0.4762	■	0.0455	0.1304
	0.0500	0.0455	■	0.1818
	0.3462	0.1304	0.1818	■

Figure 4-1: A sampling of images from the amusement park category and their confusion rates with each other.

CR of an image i is the unweighted average of its CR with each other image in its category, or:

$$\sum_{j\neq i} C_{i,j}/(N-1)$$

where $C_{i,j}$ is the confusion rate between images i and j, and N is the total number of images in the same category as i, including i itself.

Each image has an equal chance of appearing in every position, so confusions from participants misremembering an image's location (as opposed to its contents) will appear as noise and not as a bias in the CR of any particular image pair.

4.2 Generating Similarity Scores

Similarity score (**SS**) is calculated based on the choices made by participants in the similarity judgment game. We want the SS of an image pair to reflect its rank in terms of how similar its images are, as judged by participants. Participants may disagree on which of two pairs is more similar or may create a loop of judgments where each pair is judged as more similar than the one before it but less similar than the one after it. This would make it impossible to assign rankings if we attempt to make the rankings match every judgment. Thus, we formulate the ranking problem as a least squares problem so we can assign rankings even with contradictory judgments [10].

We treat each judgment as a constraint on the rankings of two pairs telling us that we must assign a larger similarity score to one vector than the other. However, the experiment does not provide us with the magnitude of the difference between their similarity scores, so we treat each judgment as providing the same amount of information and use a constant for the magnitude of each difference.

This means we are finding the vector of similarity scores x satisfying the optimization problem:

$$\arg\min_x \textstyle\sum_{i,j,k \in \mathcal{S}} (x_i - x_j - b_{i,j,k})^2.$$

where $\mathcal{S}$ is the set of all judgments made on image pairs and x_i is the similarity score for the i^{th} pair of images. Since the same two pairs can be judged multiple times, we use $b_{i,j,k}$ to represent the result of the k^{th} judgment between the i^{th} and j^{th} image pairs. It has a value of 1 if pair i was judged as more similar, and a value of -1 if pair j was more similar. In effect, we are finding the similarity scores that minimize the squared error against the constraints generated from our judgments.

If two pairs in a judgment are far apart in similarity, our similarity scores will still reflect that even though all values in b are 1. If a pair has high similarity, then there will be many judgments where that pair has higher similarity, increasing the similarity score we produce. Similarly, a pair with low similarity will have many judgments decreasing its similarity score. Thus, two pairs that are far apart in similarity will also have scores that are far apart, not just because of the judgment that directly compares them, but because of the judgments comparing each of them with other pairs.

4.3 Measuring Human Consistency

Split-subject consistency for the sequential AMT experiment as described in [13] is measured by first splitting participants into two groups at random. For each group, memorability scores for each image are calculated using the experimental results of that group only. Finally, the two sets of scores are correlated using Spearman's rank correlation, and the strength of the correlation is reported as the consistency.

Split-subject consistency for our experiments is measured similarly. Participants are split into two groups at random, the relevant statistic - CR or SS - is calculated for each group, and the two sets of results are correlated using Spearman's rank correlation.

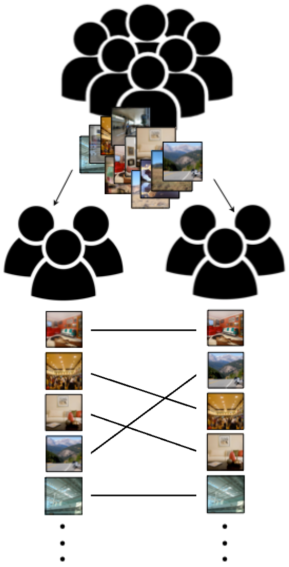


Figure 4-2: Split-subject consistency involves splitting participants into two groups, ranking the stimuli based on the results of each group, and correlating the rankings.

4.4 Building a Neural Network to Predict Confusions

We constructed our neural nets by modifying a convolutional neural net (CNN) originally trained on the ImageNet dataset. This CNN includes eight layers with weights; the first five are convolutional layers while the last three are fully-connected layers, also known as inner product layers. It also includes a number of rectified-linear, normalization, pooling, and dropout layers, which do not have weights. The last fully-connected layer's output is a 1000-length vector, which is then fed through a softmax layer to produce a distribution over 1000 classification categories.

The dataset consists of 256×256 images, but it is augmented to help prevent overfitting. Randomly sampled 227×227 patches of these images and their horizontal reflections are used as input. At test time, five patches from fixed locations and their reflections are individually passed through the CNN to produce ten predicted labels, and these are averaged to produce the final output. The structure of the CNN and the dataset augmentation are discussed in more detail in [18].

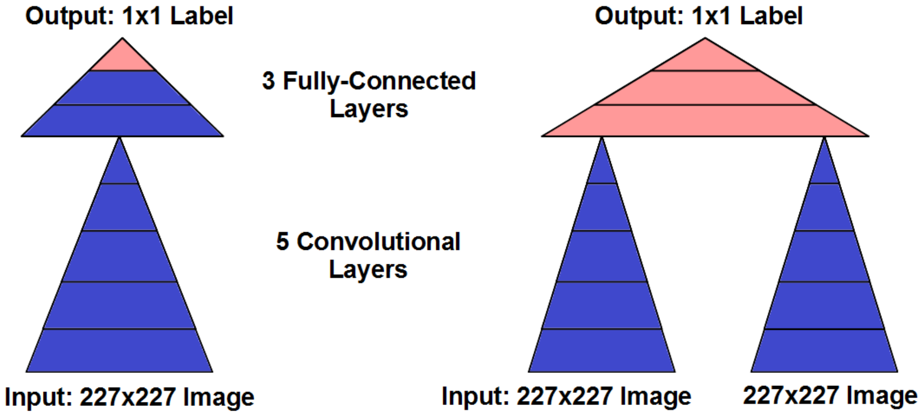


Figure 4-3: Simplified diagram of the single-input (left) and siamese (right) neural nets. Blue (dark) layers copy weights from the ImageNet CNN and are held constant. Red (light) layers are fine-tuned during training. The size of the layers is not representative of their input and output dimensions.

We use two different types of neural net, both implemented using Caffe [14]. The first is a single-input neural net used to predict statistics based on the contents of a single image: HR and FAR. The structure is almost an exact copy of the ImageNet CNN, but as we are predicting a numerical label rather than performing a classification task, we change the last fully-connected layer's output size from 1000 to 1 and remove the softmax layer. When training, the neural net's initial weights for all layers except the last fully-connected layer are copied from a pre-trained ImageNet CNN, and the copied weights are held constant. This means only the last layer has its weights affected by training.

The second type of neural net we use is a siamese neural net, which takes two images as input and predicts statistics based on pairs of images: CR and SS. The literature has examples of siamese [7] and triplet [27, 8] networks, but these networks have been used for labels explicitly based on image similarity [7, 27]. We use a similar structure because of their past success, but use it for a different set of labels.

To construct our siamese network, we make a second copy of the convolutional layers of the ImageNet CNN. Each grouping of convolutional layers takes one of the images as input. We concatenate their outputs to form the input of the first fully-connected layer, which has its input dimensions adjusted to accommodate this change. We also change the last fully-connected layer's output to 1 and remove the softmax layer, just as we did for the single-input neural net. When training, we initialize all convolutional layers with weights copied from the pre-trained ImageNet CNN and hold them constant. Thus, we only train the three fully-connected layers.

For both neural nets, we perform the same data augmentation as described above. For the siamese neural net at test time, we create the five patches and their reflections for each image to make ten pairs of inputs, where each input pair uses two patches that are from the same location and either both reflected or both not reflected.

Chapter 5

Results and Discussion

5.1 Results from Data

5.1.1 Comparison of Confusions with Memorability and Similarity

We compare the overall CR of our images with their hit rates and false alarm rates as found by [13]. For each of our image categories, we create lists containing the HR, FAR, and overall CR for each image in that category, and correlate those lists against each other.

Category	corr(CR, HR)	corr(CR, FAR)
airport terminal	-0.3583	0.3985
amusement park	-0.5622	0.3156
badlands	-0.4609	0.6217
cockpit	-0.4506	0.4917

Table 5.1: Overall confusion rate displays strong Spearman's rank correlation with hit rate and false alarm rate for each category. This shows that memorable images are less likely to be confused, and that false alarms are related to confusions between individual images. All correlations are significant at the $p < 0.01$ level.

The overall CR displays a negative correlation with HR and positive correlation with FAR, with strength varying based on category. On a general level, images with high FAR are more likely to be confused for other images, translating to a higher

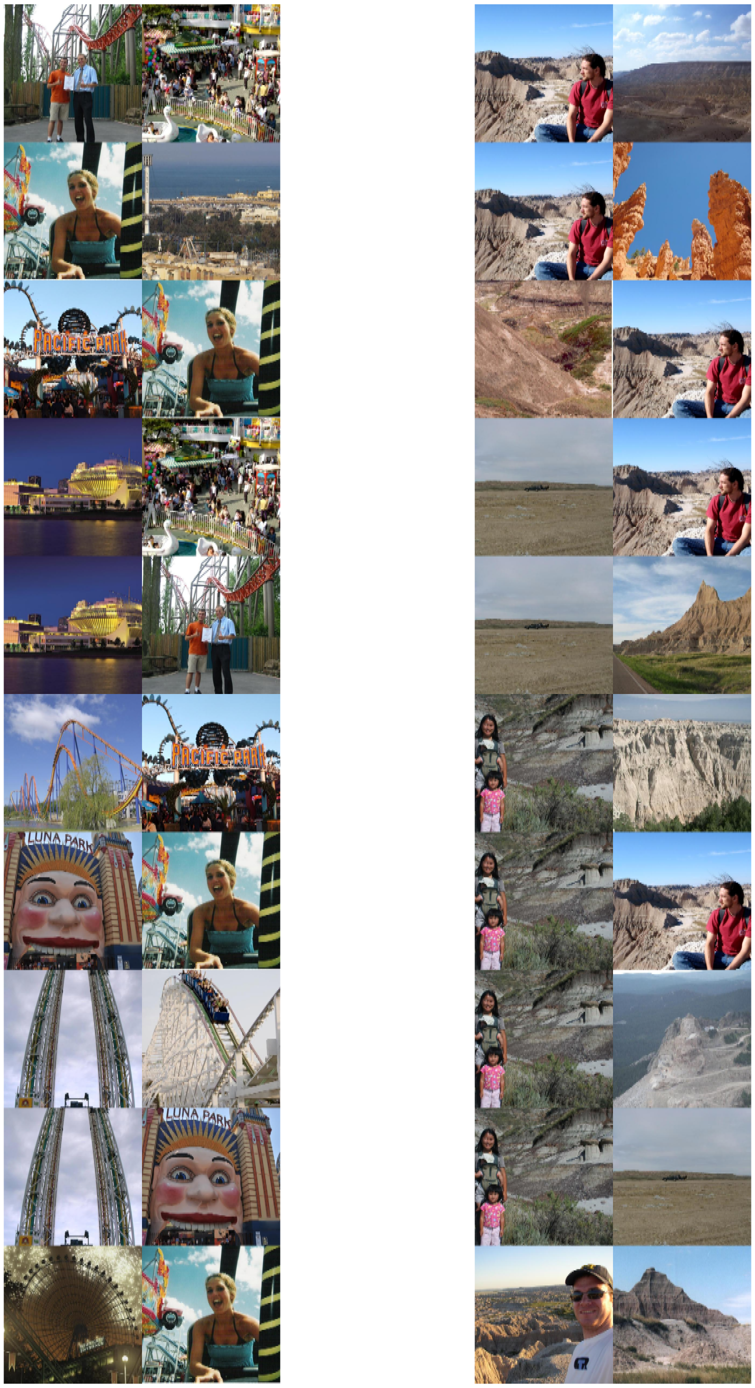
CR in the memory card game. Additionally, we see a mirror effect in which highly memorable images are correctly recognized as repeated images when they are indeed repeated, and also correctly identified as new if they are new [9, 26]. This means that HR and FAR are anti-correlated, explaining the negative correlation between CR and HR.

There are multiple hypothesized causes for false alarms. One proposes that false alarms happen if an image displays a scene with characteristics shared by a large number of previous images, even if that image is not actually similar to any of those previous images. Another proposed cause is that images with high FAR are generic and familiar, causing participants to "recognize" the image without confusing it for another image. A third possibility is that false alarms are caused by confusions between individual images, and this hypothesis is supported by our discovery of a positive correlation between CR and FAR.

Examining the least frequently confused image pairs, like those provided in fig. 5-1, show that many of them include images that contain people. Since those images tend to have high memorability [13], this means that image pairs including a memorable image are rarely confused, supporting the correlations. We also note that the HR and FAR are based on a task that included images from multiple categories, whereas CR is based on single-category tasks, making our correlations weaker than they would be otherwise.

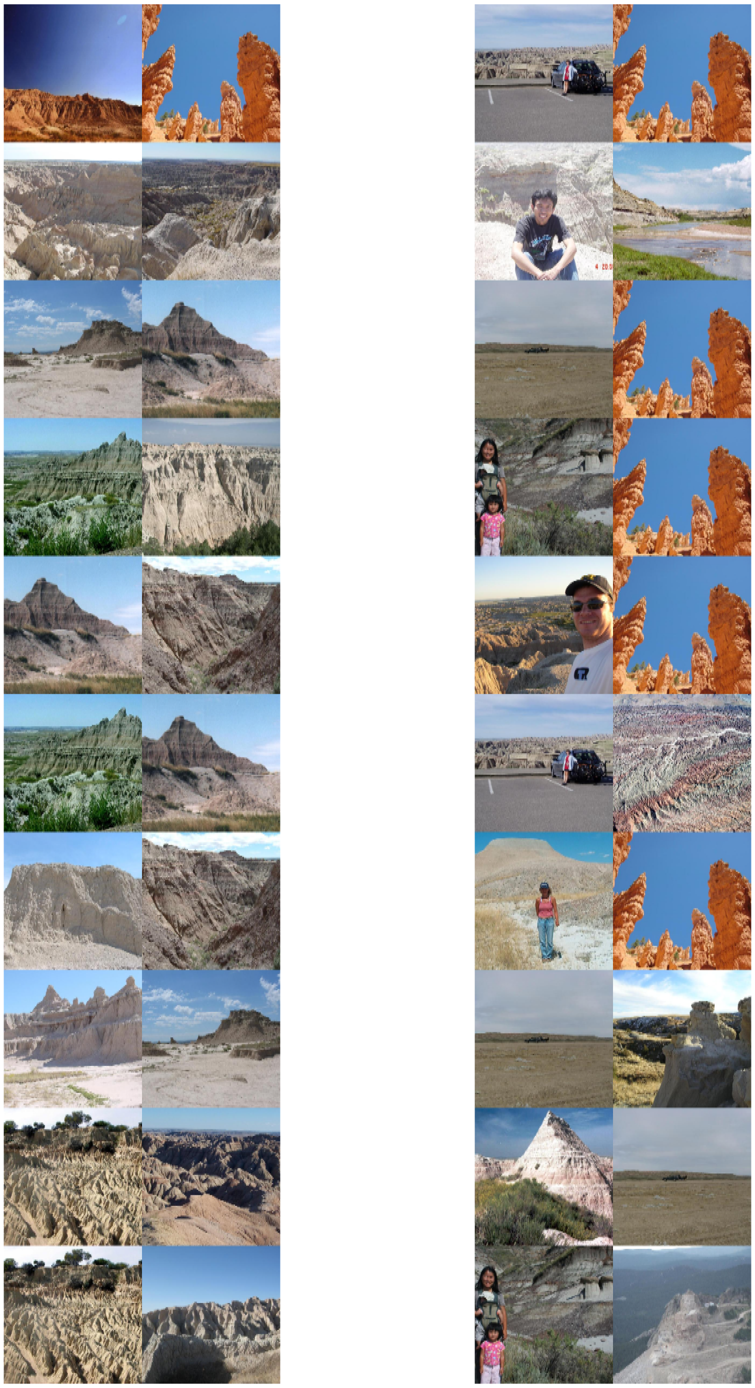
We also compare CR from the memory card game with SS from the similarity judgment task. For this, we rank the image pairs by CR and SS, separately, and then correlate the ranked lists of image pairs.

The correlation varies by category, and it is worth noting that the correlation of CR with SS shares the same weakest-to-strongest ordering as the correlation of CR with FAR. Additionally, when comparing the least and most confused image pairs (appendix D) with the least and most similar pairs (appendix E), we see that all categories have some overlapping images between the two sets of pairs, further supporting this correlation.



(a) *amusement park* (b) *badlands*

Figure 5-1: The least confused image pairs for the *amusement park* and *badlands* categories. These images contain people and faces, features associated with high memorability. Most and least confused pairs for all categories are shown in appendix D
.



(a) Most similar pairs

(b) Least similar pairs

Figure 5-2: The most and least similar image pairs for the *badlands* category, provided as an example. Note that some images appearing in the least similar images here also appear in fig. 5-1. Most and least similar pairs for all categories are shown in appendix E

.

Category	corr(CR, SS)
airport terminal	0.2416
amusement park	0.1493
badlands	0.4508
cockpit	0.3716

Table 5.2: Confusion rate displays correlation with similarity score via Spearman's rank correlation. This means that people are more likely to confuse the pairs of images that they find similar. The correlation's strength varies by category for an unknown reason.

5.1.2 Split-Subject Consistency

In order to reduce the effect of random chance when dividing the subjects into two groups, we compute split-subject consistency over 25 random splits and average the results.

Category	CR Consistency	Std. Deviation	Mean appearances per pair
airport terminal	0.0869	0.0101	21.11
amusement park	0.1213	0.0147	25.61
badlands	0.3696	0.0182	33.98
cockpit	0.2266	0.0130	21.89

Table 5.3: Split-subject consistency for CR for each category, averaged over 25 random splits. The low consistency is caused in part by the low number of appearances per image pair; consistency is likely to increase as we gather more data.

We have roughly equal amounts of data for each category; the variation in mean appearances per image is caused by differing numbers of images in each category. The *airport terminal* category contains the most images, while *badlands* contains the fewest.

These consistency values are lower than the consistency of 0.75 found in [13]. However, our results include much fewer mean appearances per image pair than [13], which has 80. Additionally, fig. 5-3 shows that consistency continues to increase with mean appearances, and has not yet begun to plateau. This suggests that the low consistency is primarily caused by an insufficient quantity of data at this point.

Additionally, the mean CR, averaged across all image pairs, is less than 20% for

Category	SS Consistency	Std. Deviation	Mean appearances per pair
airport terminal	0.3532	0.0145	18.49
amusement park	0.5734	0.0101	23.66
badlands	0.6092	0.0102	31.71
cockpit	0.5173	0.0090	20.61

Table 5.4: Split-subject consistency of similarity score for each category, averaged over 25 random splits. The consistency is higher than for confusion rate because the similarity judgment task generates a data point for every pair of images that appears, whereas the memory card game only generates a data point when a confusion occurs, and confusions occur infrequently.

each category, so many image pairs are confused only a few times. Consistency for CR is dependent on having larger numbers of confusions, but the low CR means that a low proportion of appearances for a pair actually generate a useful data point. In contrast, for the similarity judgment task, the consistency measure is affected regardless of whether an image pair is judged as more or less similar, meaning that every appearance of an image pair generates a data point. This explains why SS consistency is higher than CR consistency despite having fewer mean appearances per image pair; the similarity judgment task generates more data within those appearances.

However, observe that in fig. 5-3, the CR consistency for *badlands* passes the values given for other categories in table 5.3 even before reaching 20 mean appearances per image. This means that insufficient data is not the sole cause of the low consistency, and that some property of the category can cause high or low consistency measures. One such property could be how similar the images are within a category.

In light of these consistency values, the correlations shown in table 5.2 may be stronger than they previously appeared. We measured Spearman's rank correlation between CR, calculated from half of the subjects, and SS, calculated from the other half of the subjects. We took the mean of this correlation over 25 trials and compared this value to the geometric mean of the split-subject CR consistency and the split-subject SS consistency, which we use as an upper bound for the split CR-SS correlation [22]. The results are shown in table 5.5.

We find that the correlation between CR and SS is comparable to the internal

Category	corr(CR, SS)	Upper bound
airport terminal	0.1176	0.1752
amusement park	0.0886	0.2637
badlands	0.3343	0.4745
cockpit	0.2308	0.3424

Table 5.5: The correlation is between confusion rate based on half the subjects selected at random, and similarity score based on half the subjects selected independently and at random, averaged over 25 trials. The estimated upper bound is the geometric mean of confusion rate consistency and similarity score consistency for that category. The correlation between confusion rate and similarity score is close to the internal consistency of the two measures, suggesting that the correlation is stronger than the values in table 5.2 would indicate.

consistency for the two measurements. This implies that the relationship between CR and SS is stronger than the values in table 5.2 would suggest, and that this correlation is likely to increase further as we gather more data and increase our split-subject consistency. Note, however, that the CR-SS correlation does not account for all of the explainable variance, so there are still differences between how similarity is judged and how images are encoded in memory space.

5.2 Results from Neural Nets

5.2.1 Neural Nets for Predicting Memorability

We first use the single-input CNN structure as discussed in section 4.4 to establish a baseline for the ability of CNNs to predict memorability scores. For these results, the dataset consisted of all images from [13] for which HR and FAR data were available, split into a training and test set at random. We trained one network to predict an image's HR, and one to predict an image's FAR. The HR and FAR scores were converted to integer labels within the range 0-100. Both networks were trained for 30,000 iterations using the squared difference between the predicted and expected score as the training error. The training parameters are provided in greater detail in appendix A.

5.2.2 Neural Nets for Predicting Image Confusions and Semantic Similarity

For tasks involving the siamese neural net, we first split the images in each category into a training group and a test group, at random. Then for each group, we create all possible pairings of images within the group, as shown in fig. 5-5. The set of pairings constructed from the training group becomes the training set, and the set of pairings from the test group becomes the test set. Each pair appears twice in the training or test set, but with the order of its images swapped, in order to encourage symmetric layer weights.

We split on images so the neural net operates on entirely unfamiliar images at test time. If we had split on pairs, the neural net could overfit by associating specific images with high or low labels and guessing the label based on that information. Rather, our model should be able to make predictions about a pair of images without having previously seen those images.

We trained three CNNs with this network structure. Two of them simply use the training and test sets as described above; one is trained using the CR for each pair, while the other uses the SS. The third CNN is trained on CR, but uses an augmented training set. For each image in the training group, we construct a pair where that image occurs twice and assign it a CR of 100%.

Adding the pairs of duplicates effectively enforces the constraint that the CNN should assign a high label to very similar images. Recall that the number of confusions is low, meaning that the training and test set are affected by noise. Adding the pairs of duplicates adds reliable information to the training set.

CR, like HR and FAR, is converted to an integer label from 0-100. SS values first had a constant added to make them all nonnegative, then multiplied by 100 and rounded. This causes the SS labeling to have a different range than the CR labeling, but in practice, scaling the labels did not affect the predictive strength of the CNNs.

We observe that the siamese neural net structure is capable of predicting confusion rates, especially when the data is made more reliable with the addition of duplicate

Predicted Statistic	Pearson corr	Spearman's rank corr
Confusion rate	0.09094	0.09096
Confusion rate with duplicates	0.17134	0.22836
Similarity score	0.24837	0.22619

Table 5.6: Predictive strength on the test set for each neural network, as measured by correlation between the expected and CNN-predicted values. Duplicates refer to the pairs with two matching images. Scatterplots of predicted and actual scores are available in appendix G.

pairs. This confirms that confusions are a modelable process, and that confusions between images can be predicted from the contents of the images themselves.

The CNN's superior performance on SS can be attributed to the higher consistency of SS data and therefore higher reliability of the training set. It may also be because SS is based solely on the properties of images, while CR can be affected by extrinsic factors like noise and context.

5.2.3 Neural Nets and Consistency

We trained three more CNNs while combining the split-subject test for consistency with the training-test split on images. In addition to creating the training and test sets, we split the subjects into two groups at random. We then use the data from one subject group to create the training set labels, and the data from the other subject group for the test set labels, as shown in fig. 5-6. This is the closest reflection of the ideal use of a predictive model: using data to make predictions on new images seen by future users. Again, each pair appears twice in the training or test set, once with the order of its images swapped.

Our results are listed in table 5.7.

The low consistency for CR translates to non-significant prediction strength in the split-subject neural networks. The prediction strength and consistency for CR may be caused by context. The set of images in each game is completely randomized, so a given image pair will not appear in the same context each time. As a result, the "true" confusion rate of an image pair may vary between subject groups simply

Consistency	Pearson corr	Spearman's rank corr
Confusion rate	NS	NS
Confusion rate with duplicates	0.06592	NS
Similarity score	0.18285	0.17452

Table 5.7: Split-image-and-subject consistency for each neural network, as measured by correlation between the expected and CNN-predicted values. Scores are omitted if they are not significant at the $p < 0.01$ level. Consistency here is affected by the consistency of the original data. Scatterplots of predicted and actual scores are available in appendix H.

because they see the image in different contexts.

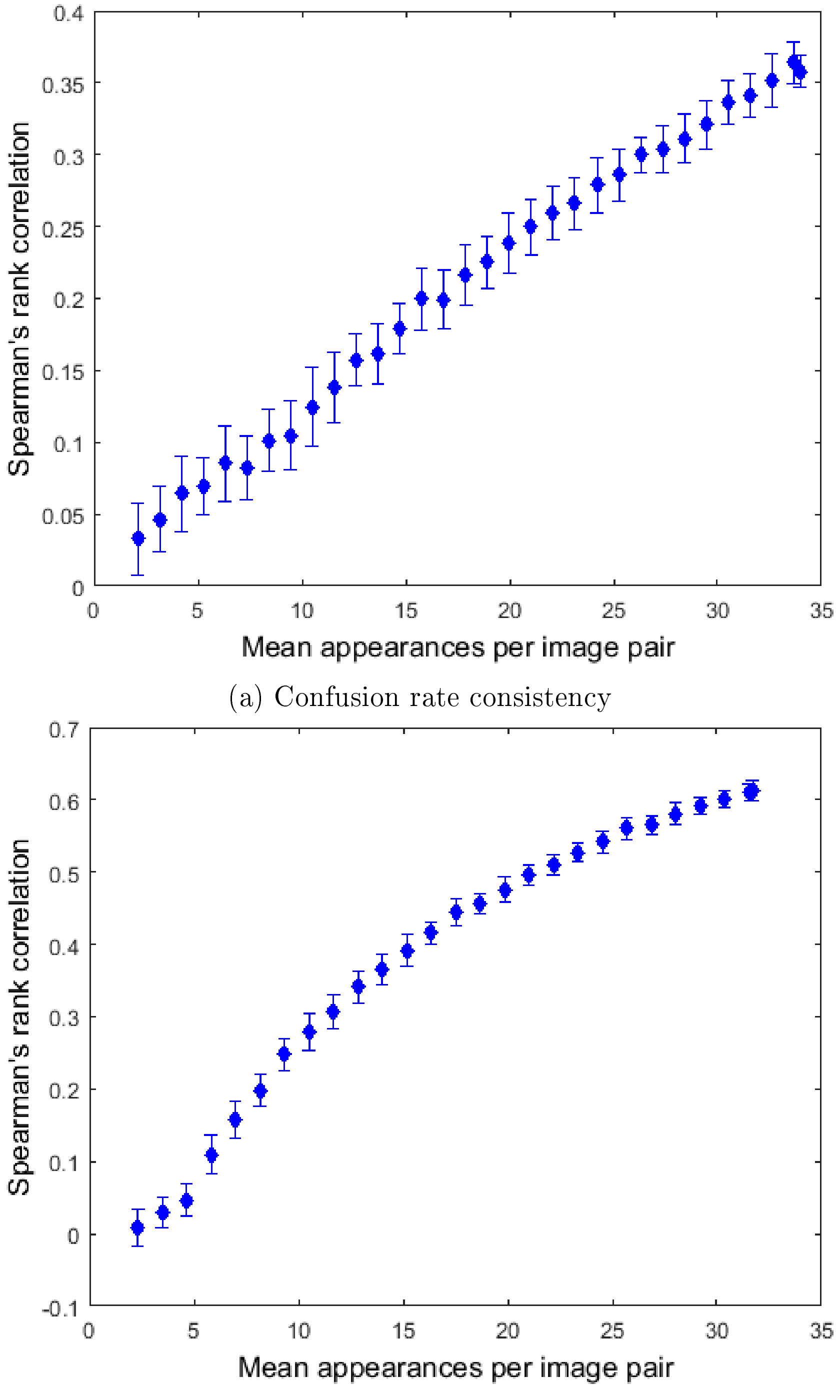
Meanwhile, because SS had high consistency, its neural net retains much of its predictive strength. Just like with split-subject consistency, the predictive strength of these CNNs is likely to increase as we gather more data.

5.3 Key Findings

We discovered a correlation between CR and FAR, which suggests that false alarms may be caused at least in part by confusions between individual images. This correlation also indicates that our memory card game is a valid experimental task, and that it is not producing arbitrary and inaccurate data.

We discovered a correlation between CR and SS, which indicates that people are more likely to confuse pairs of images that they find similar. This also suggests that an image's representation in memory space is related to its representation in visual space.

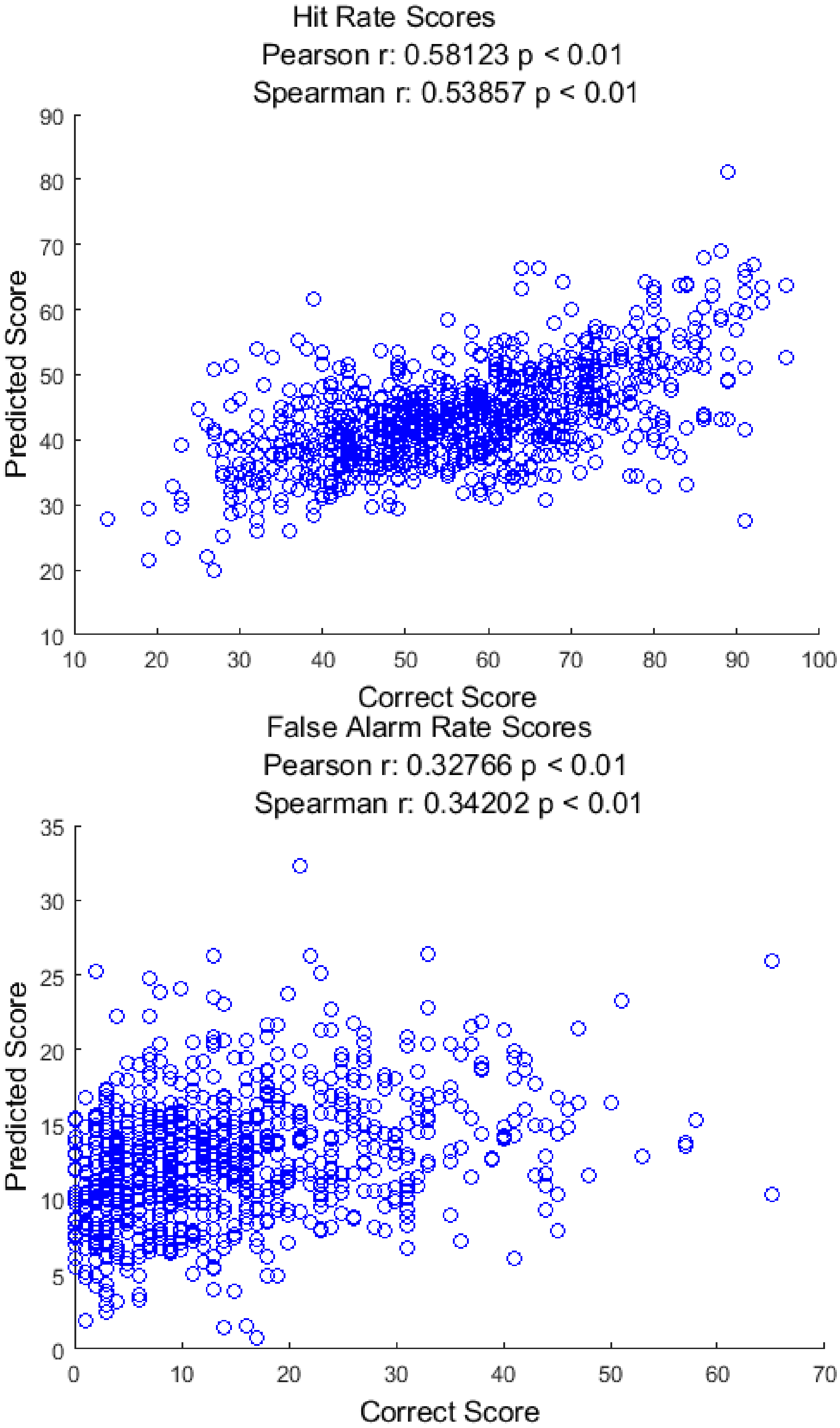
We also found that the siamese neural net is capable of predicting CR when given pairs of images as input, and the predictive strength of this neural net appears to increase with the consistency and reliability of the data it is based on.



(a) Confusion rate consistency

(b) Similarity score consistency

Figure 5-3: Split-pairs consistency for CR (top) and SS (bottom) as a function of mean appearances per image pair, for the *badlands* category. Consistency rises with mean appearances, and has not yet plateaued. Error bars show standard deviation over 25 trials. Graphs for the other categories can be found in appendix F.



Hit Rate Scores
Pearson r: 0.58123 p < 0.01
Spearman r: 0.53857 p < 0.01

False Alarm Rate Scores
Pearson r: 0.32766 p < 0.01
Spearman r: 0.34202 p < 0.01

Figure 5-4: Scatterplots of neural network predictions for hit rate (top) and false alarm rate (bottom), plotted against their actual values.

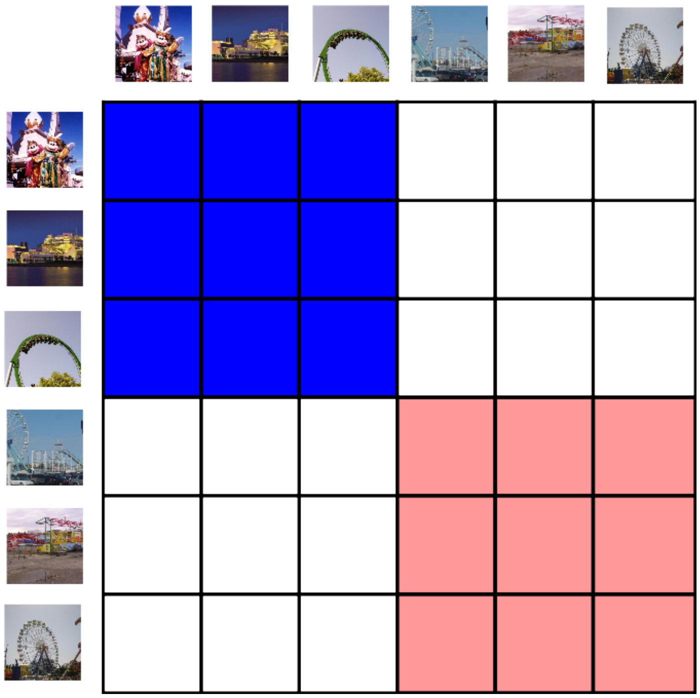


Figure 5-5: We split the images into two groups, and construct the training and test set from all image pairs within each group. The blue (dark) area is the training set, and the red (light) area is the test set.

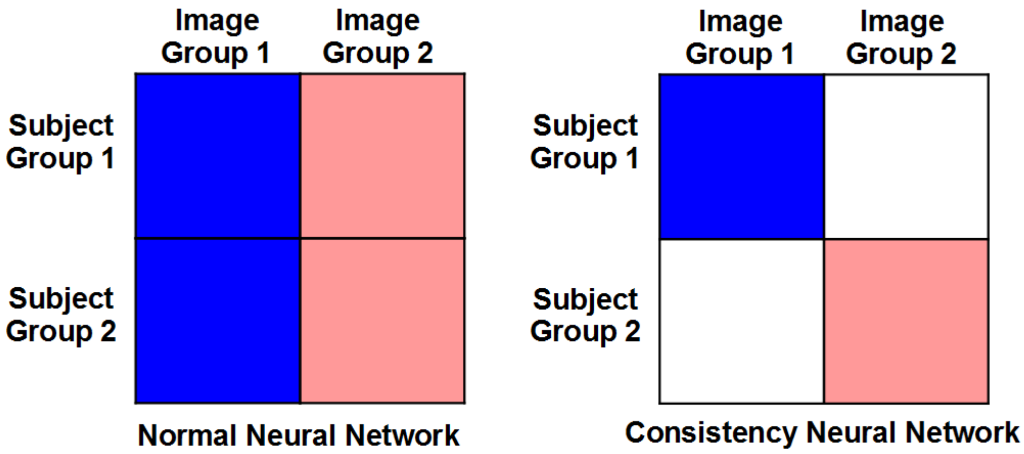


Figure 5-6: We normally split only on images, and use all subjects to create the labels for both sets. For the neural nets used to evaluate consistency, we use half the subjects and half the images for the training set and the other half of the subjects and images for the test set. The blue (dark) area is the data used for the set, and the red (light) area is used for the test set.

Chapter 6

Conclusion

In this paper we have established a new experimental task, which fills a gap in the literature by allowing us directly study failures of memory in the form of confusions between pairs of images. Additionally, we were able to show that this task generates results consistent with the literature, meaning that it is a valid experimental task and a useful tool for exploring failures of memory.

We found that memorable images tend to be rarely confused, and that image pairs with low confusion rate often have characteristics associated with high memorability, especially the presence of people in images. We also found that high false alarm rate is associated with high confusion rate, shedding light on the causes of false alarms and showing that they are related to confusions between images.

We also found that there exists a strong correlation between the images that people confuse and the ones that they find similar. Additionally, some images that appear in the most confused image pairs for each category also appear in the most similar image pairs, and the same applies for the least confused and least similar image pairs. This suggests that the mental process for remembering and recalling images overlaps with the process for evaluating the similarity of two images. At the same time, the strength of the correlation and the differences in the images appearing in those pairs indicate that visual encoding is not the same as memory encoding. In other words, people may rely on one encoding and set of features to store images in memory, and a different but related encoding and set of features for comparing freely visible images.

Finally, we have established that convolutional neural networks can be used to create models for predicting confusions between images. The predictive strength of these models tells us that confusions between images are caused at least in part by the content of those images. Furthermore, we showed that this can be done using a siamese neural net, opening up new avenues of research and predictive models that we could not explore with single-input CNNs.

6.1 Future Research

Our work lays part of the foundation for a model of memory. If we can create a CNN that accurately predicts CR for pairs of images, we can examine its layers and feature vectors to determine what parts of the image are predictive of its confusions, and hopefully determine what characteristics of an image are stored in memory. We can also examine how the feature vectors are combined to create a label, which would show us what causes images to be confused for each other, whether it is due to similar representations in memory or some other set of factors.

Additionally, we can construct custom sets of images for memory tasks, similarly to how images were generated in [20], to determine what changes can be made to an image without affecting its representation in memory.

We can also backtrack and explore properties of context. The literature has already established that context affects memorability [6, 17, 3, 11, 23], but by understanding how images are represented in memory space and what causes them to be confused, we can learn what features cause an image to blend into or stand out from its context.

Finally, a better understanding of memory and confusion is useful for any application that relies on or benefits from accurate storage and recall of information, including areas such as teaching and graphic design.

Appendix A

Neural Net Parameters

For both neural nets, the initial learning rate was 1×10^{-5}, chosen because it was the largest learning rate that would not cause the training error to diverge given the other parameters and the initial layer weights. The momentum was 0.9, and the weight decay was 5×10^{-4}.

The single-input neural nets were trained for 30,000 iterations with a batch size of 64, and we multiply the learning rate by 0.1 every 10,000 iterations.

The siamese neural nets were trained for 100,000 iterations with a batch size of 64, and we multiply the learning rate by 0.1 every 34,000 iterations.

Appendix B

Findings on Spatial Confusion

When the participant confuses a pair of images, we check if the participant has previously revealed both the second image and the matching counterpart to the first image. If so, this pair of clicks is relevant to spatial confusion. For each pair of clicks that meet these requirements, we record the displacement between the location of the match for the first image (that is, the "correct" location for the participant to have clicked on) and the location of the second image in the confusion pair.

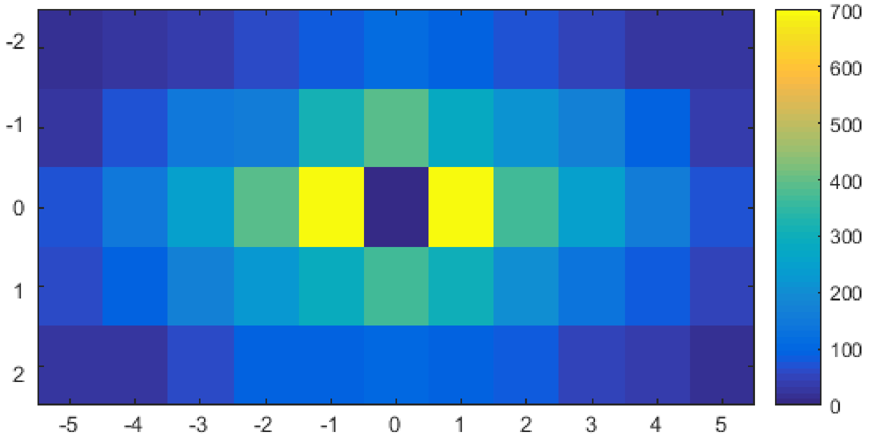


Figure B-1: Heatmap showing the frequency of spatial confusions based on the displacement between the location of the matching image and the location of the user's click. The distribution is roughly symmetric, and error frequency decreases as displacement increases.

We ran a simple check on the *airport terminal* category, counting the number of click pairs with each possible displacement. The result is shown in B-1, where $(0, 0)$ represents the location of the match for the first image. The largest number of confusions occur immediately to the left and right of the correct location, and confusion frequency drops rapidly as we move away from it. We also observe that the distribution of errors displays both vertical and horizontal symmetry.

Appendix C

Improving Neural Net Predictions

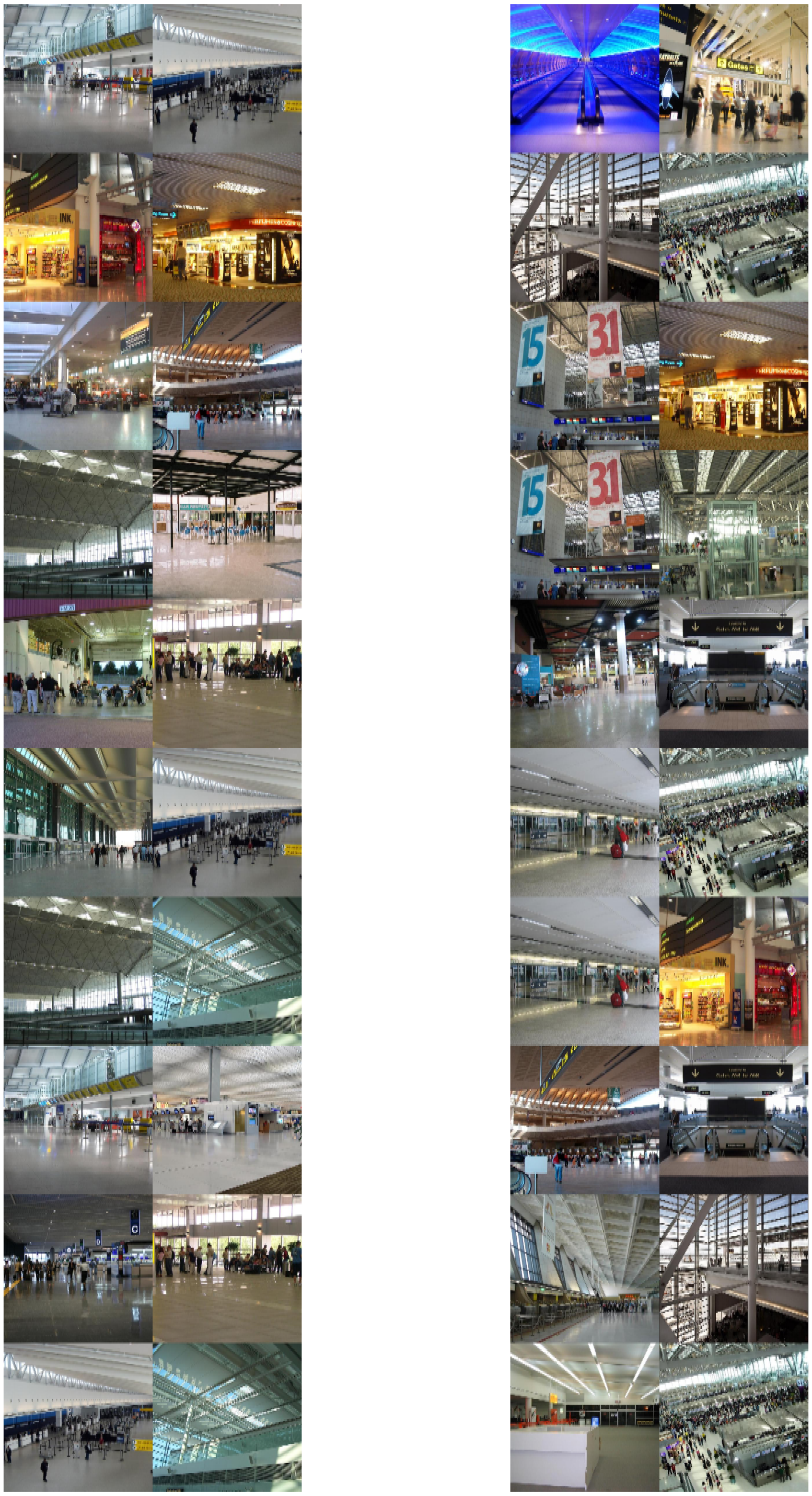
One major issue in this paper was the low consistency for confusion rate data, caused by a combination of confusions being uncommon events and low mean appearances per image pair. For future work using the memory game, the easiest way to alleviate this issue is to increase the number of distinct images appearing in a single game. This puts a greater strain on participants' memory, making confusions more frequent. Additionally, the number of image pairs appearing in a game grows quadratically with the number of distinct images, allowing each game to produce information for a larger number of image pairs without overwhelming participants with a massive number of images.

We used images from the SUN database so we could compare existing HR and FAR data with CR data generated by our task. This caused difficulties later on because the fixed weights for the convolutional layers were trained to extract features from ImageNet images, and those features are not necessarily the same as those present in SUN images. Further research with these CNN structures should use a subset of ImageNet images so the layer weights match the images they were trained for.

It is also possible to improve CNN performance by manually tuning the fully-connected layers of our CNN structure. The first layer's input dimensions and the last layer's output dimensions must remain fixed, but that leaves two sets of dimensions that can be experimented with. The output dimensions of the first and second fully-connected layers can be adjusted, and the input dimensions of the second and third fully-connected layers, respectively, should match them. It may be worth experimenting with different dimension sizes to find the layer specifications that produce the best predictors. We may also get improvement by adjusting the parameters for the randomized initial weights of these layers so that the range of early predictions more closely matches the range of expected scores.

Appendix D

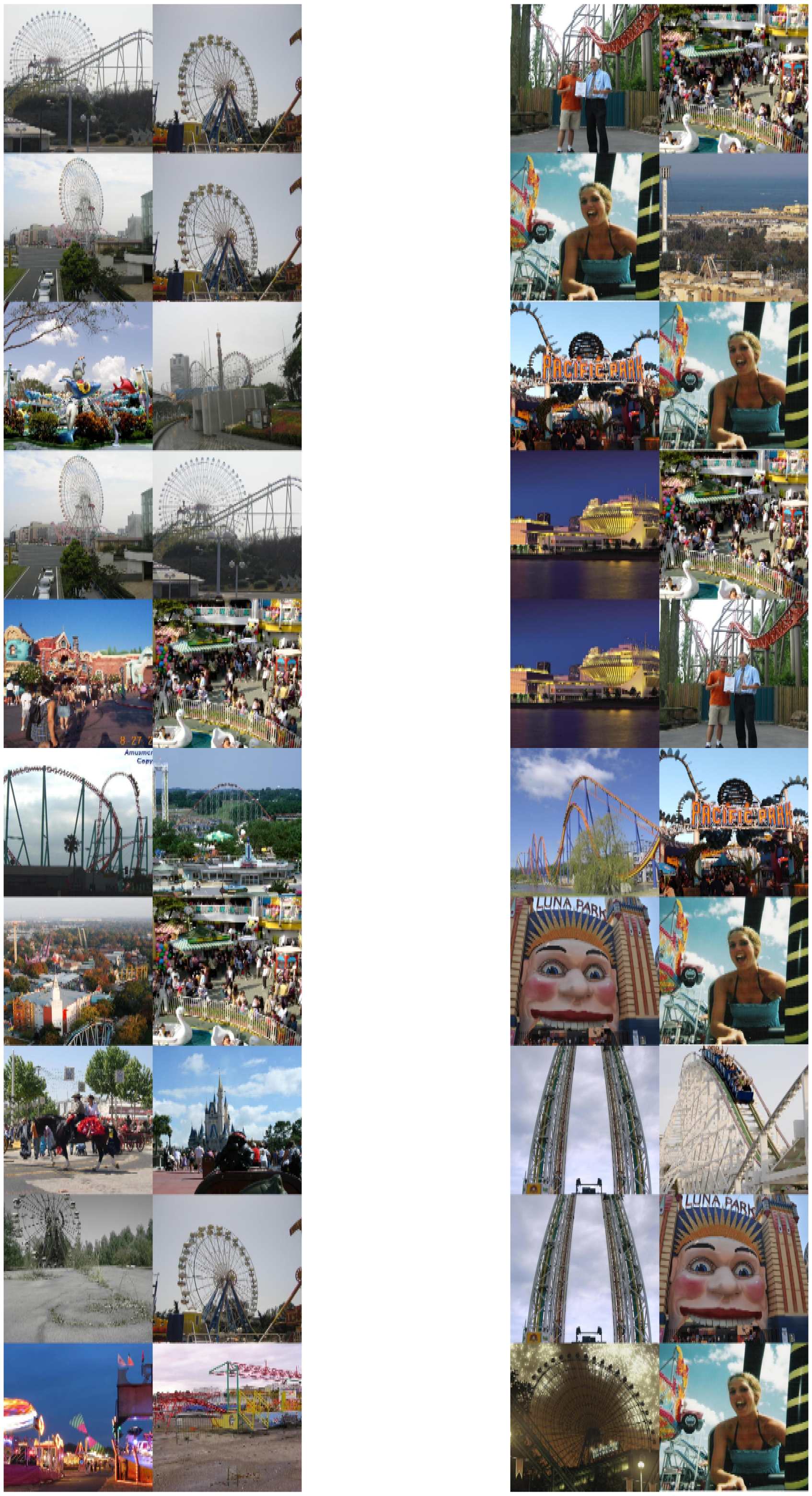
Image Pairs by Confusion Rate



(a) Most confused pairs

(b) Least confused pairs

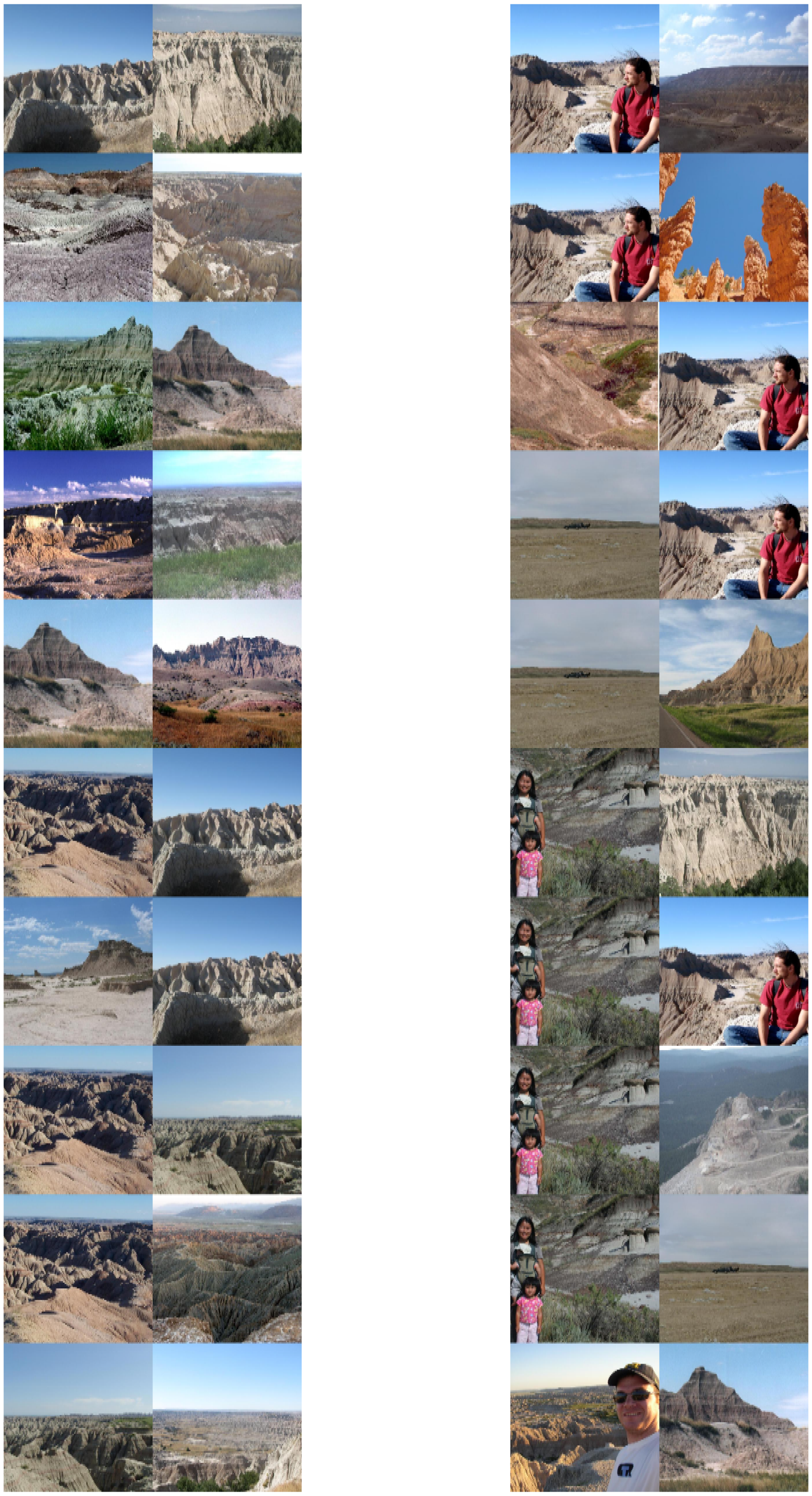
Figure D-1: Category: *airport terminal*



(a) Most confused pairs

(b) Least confused pairs

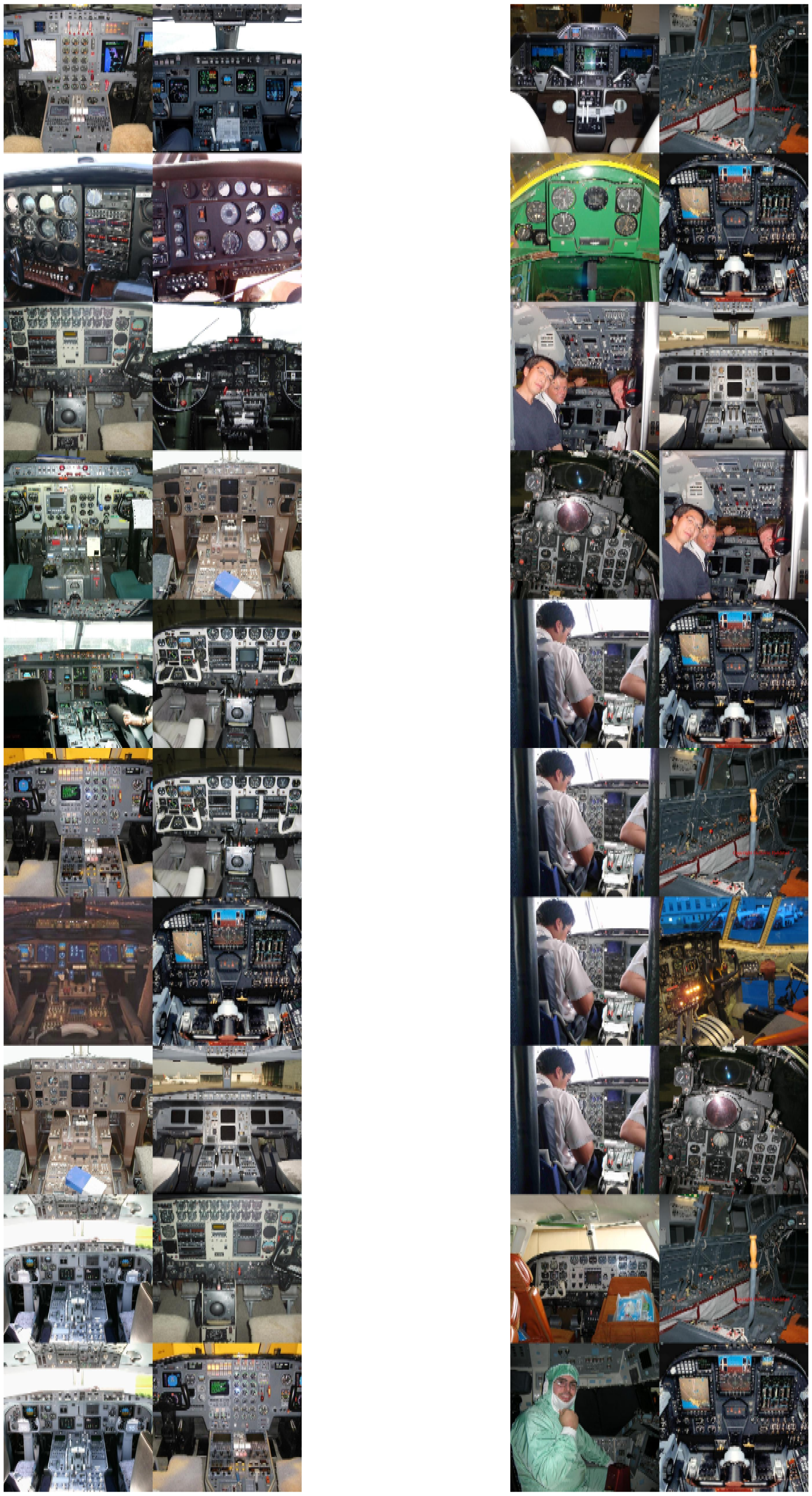
Figure D-2: Category: *amusement park*



(a) Most confused pairs

(b) Least confused pairs

Figure D-3: Category: *badlands*



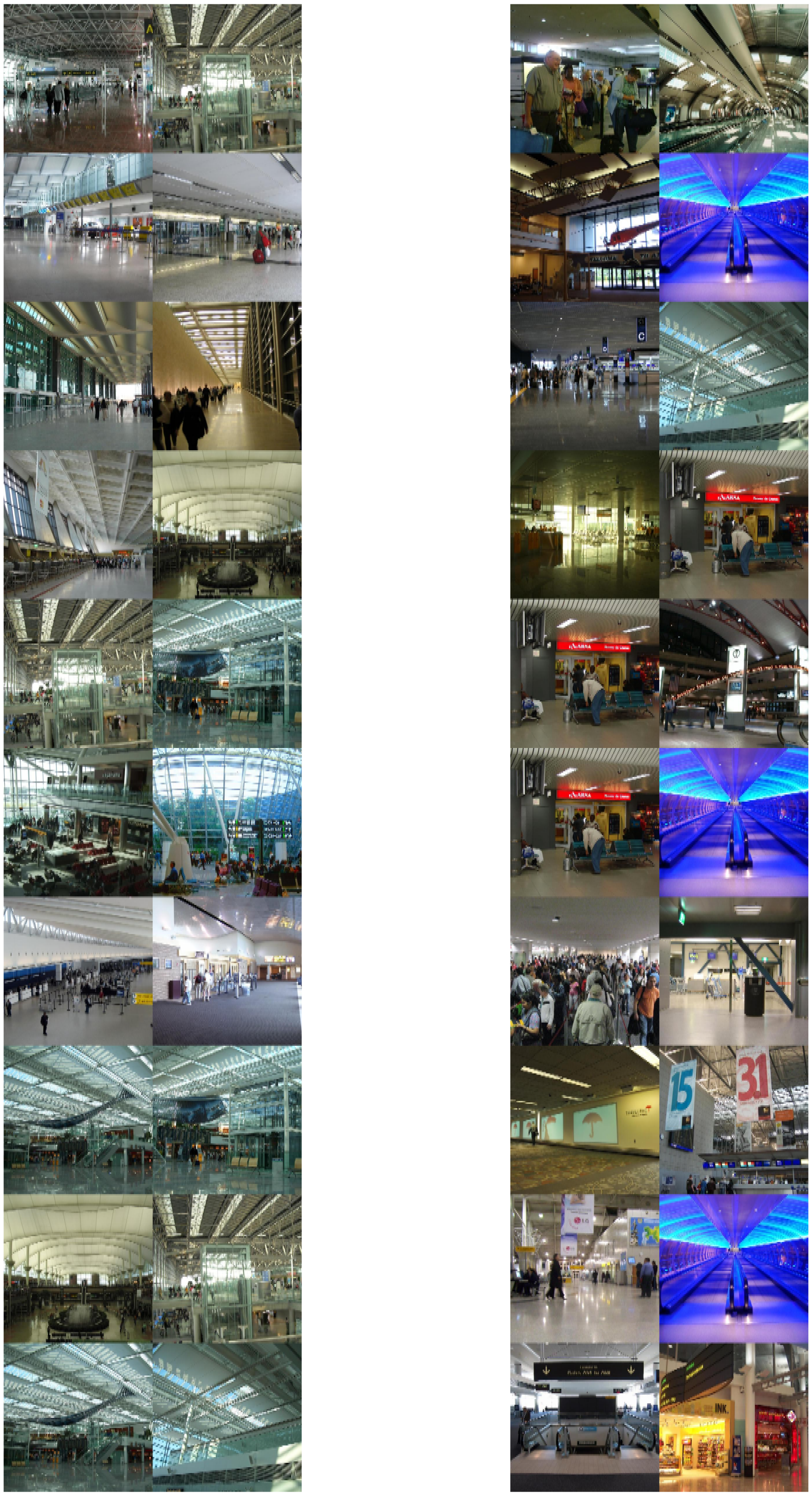
(a) Most confused pairs

(b) Least confused pairs

Figure D-4: Category: *cockpit*

Appendix E

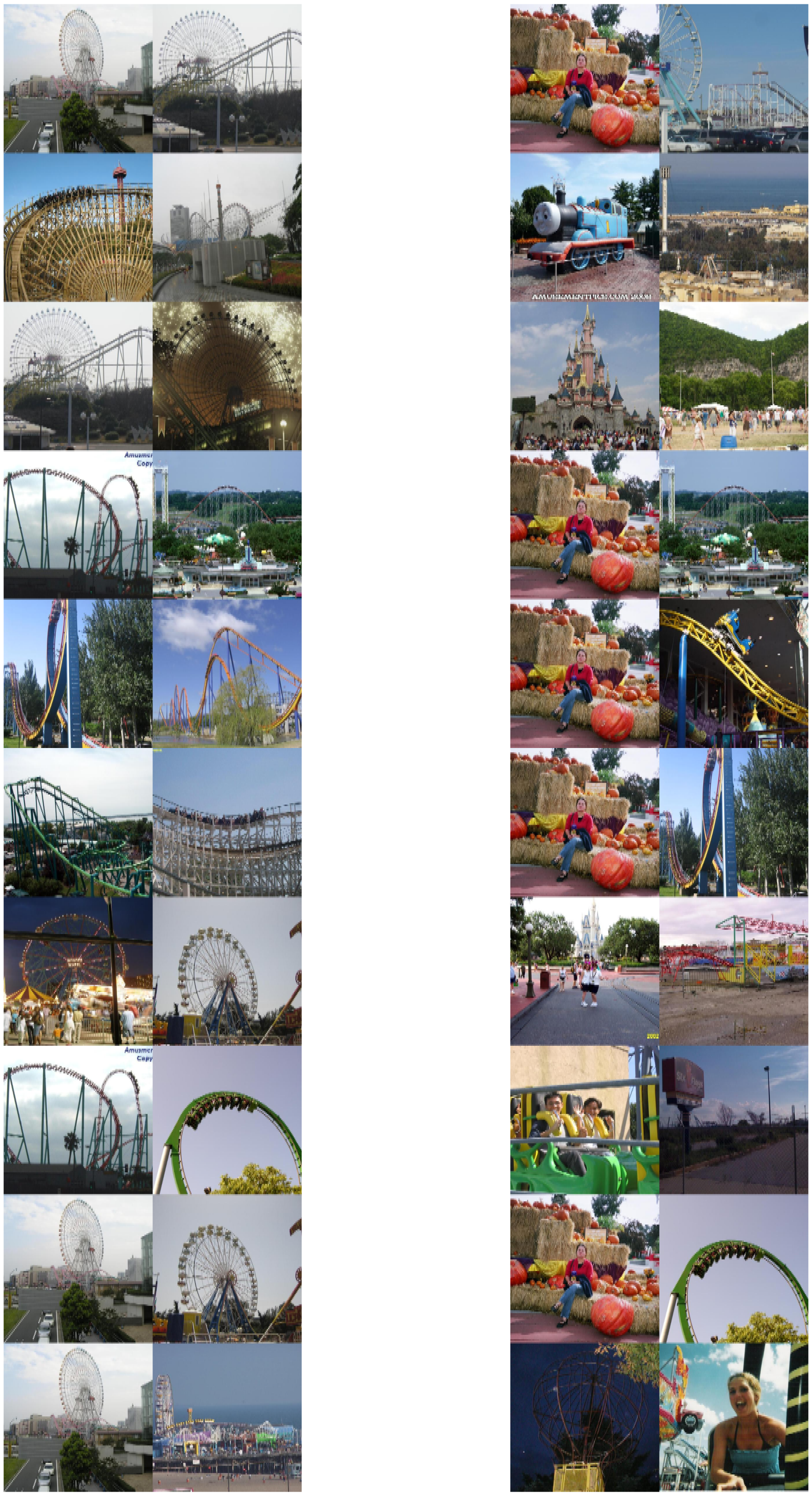
Image Pairs by Similarity



(a) Most similar pairs

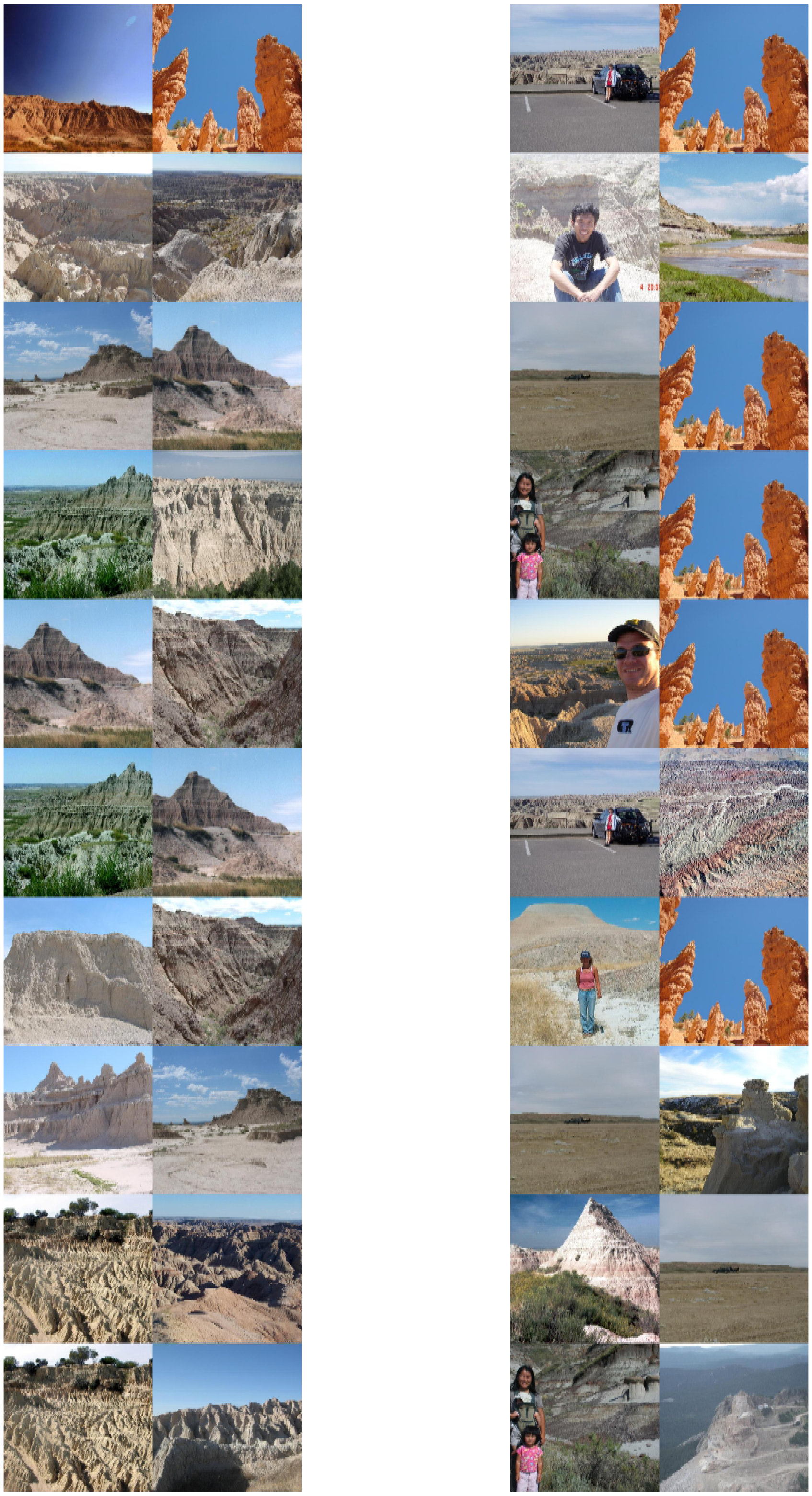
(b) Least similar pairs

Figure E-1: Category: *airport terminal*



(a) Most similar pairs (b) Least similar pairs

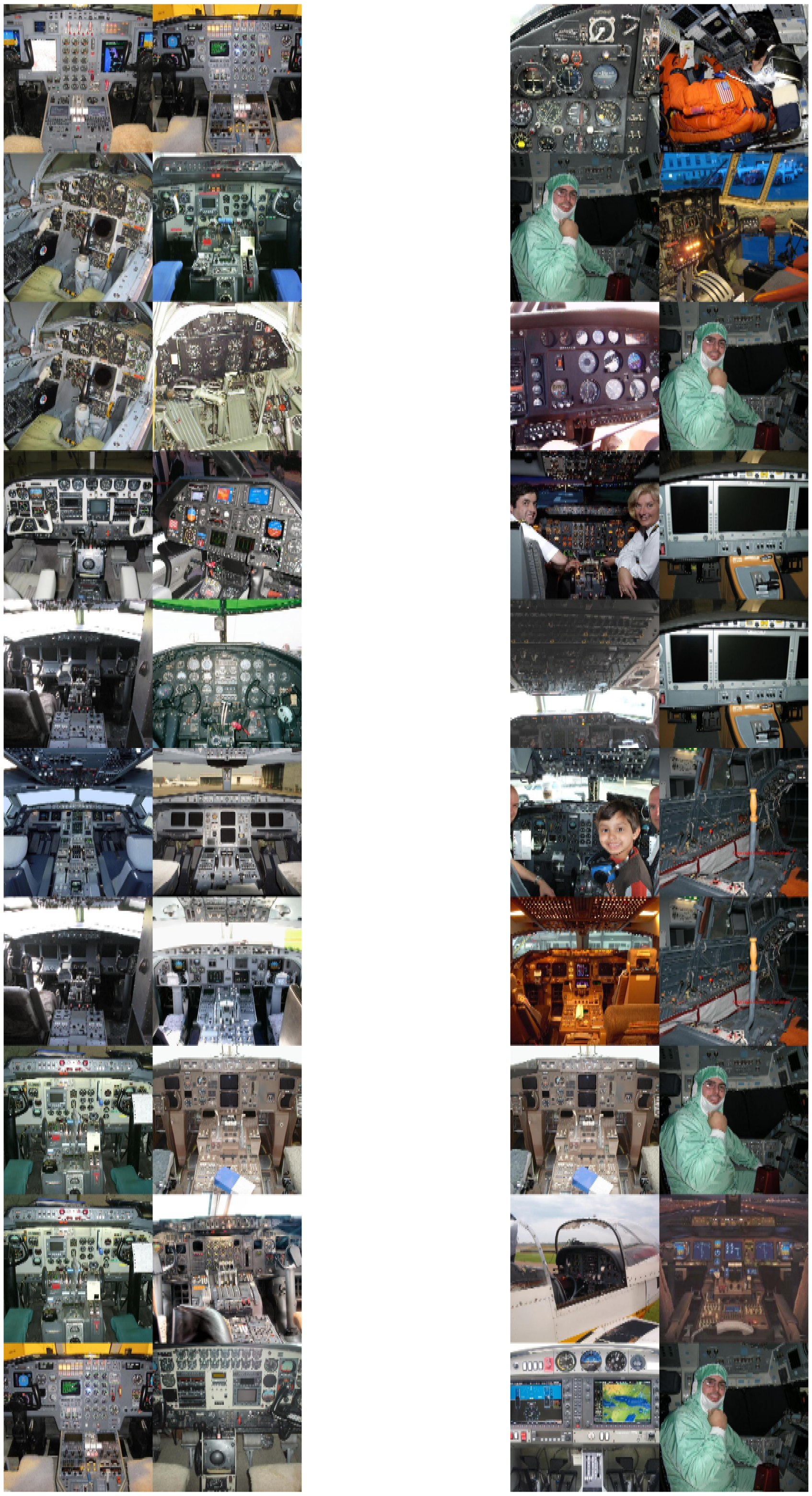
Figure E-2: Category: *amusement park*



(a) Most similar pairs

(b) Least similar pairs

Figure E-3: Category: *badlands*



(a) Most similar pairs

(b) Least similar pairs

Figure E-4: Category: *cockpit*

Appendix F

Consistency Graphs

The graphs for the badlands category are duplicated here for completeness.

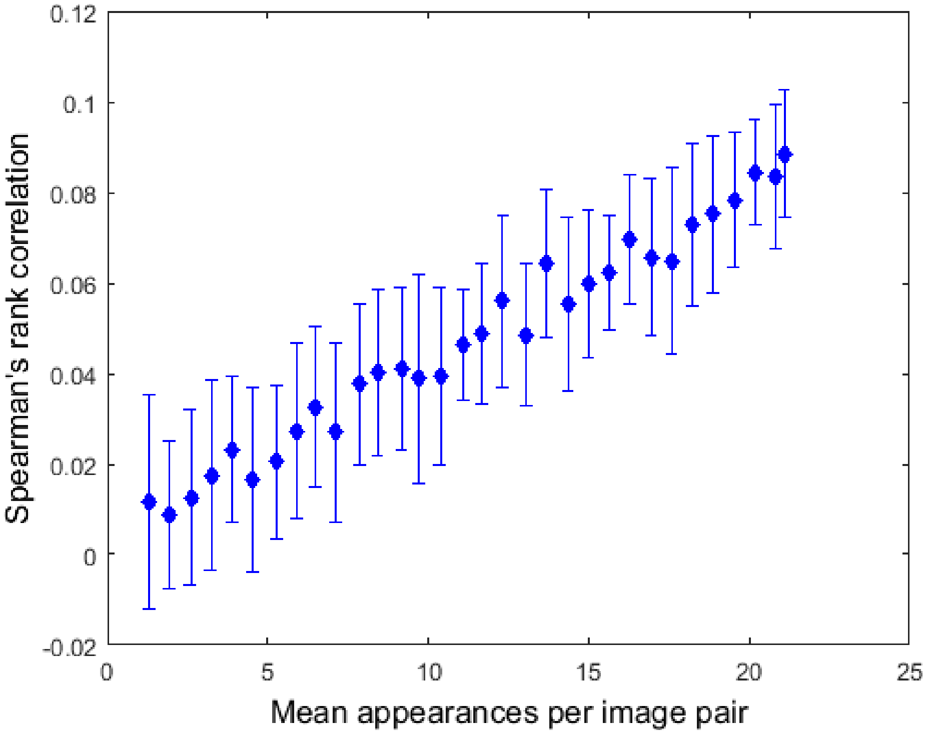


Figure F-1: *airport terminal* confusion rate consistency

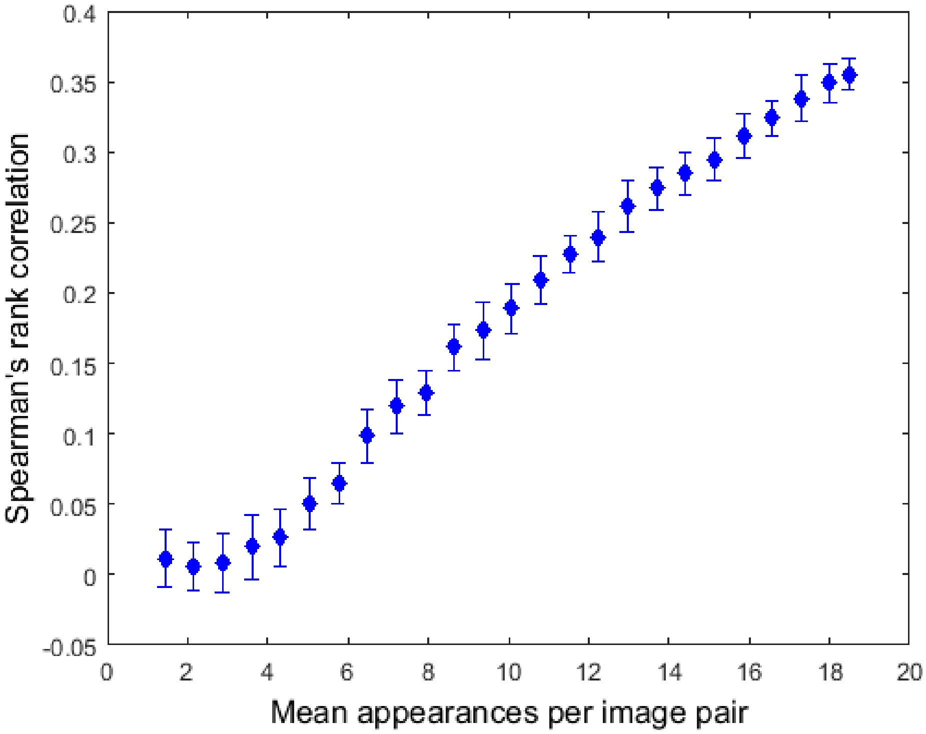


Figure F-2: *airport terminal* similarity score consistency

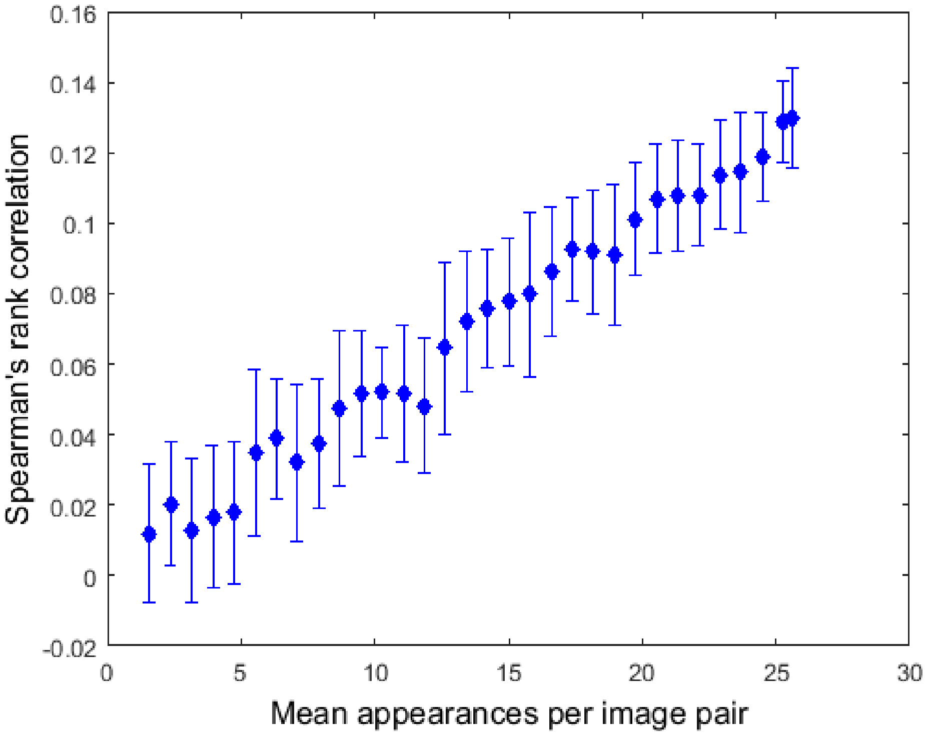


Figure F-3: *amusement park* confusion rate consistency

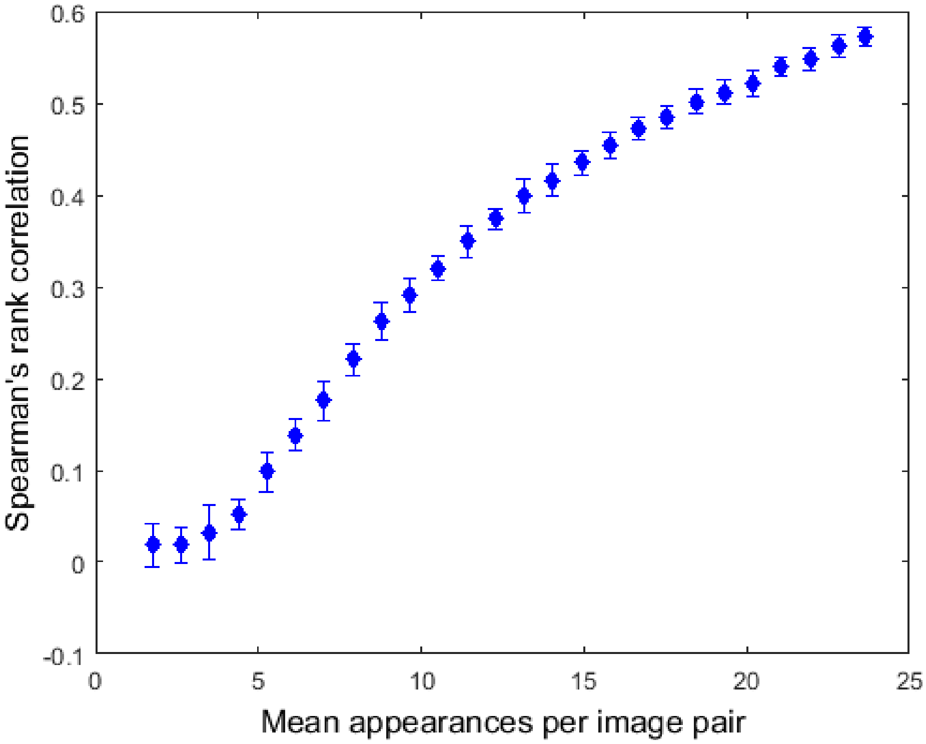


Figure F-4: *amusement park* similarity score consistency

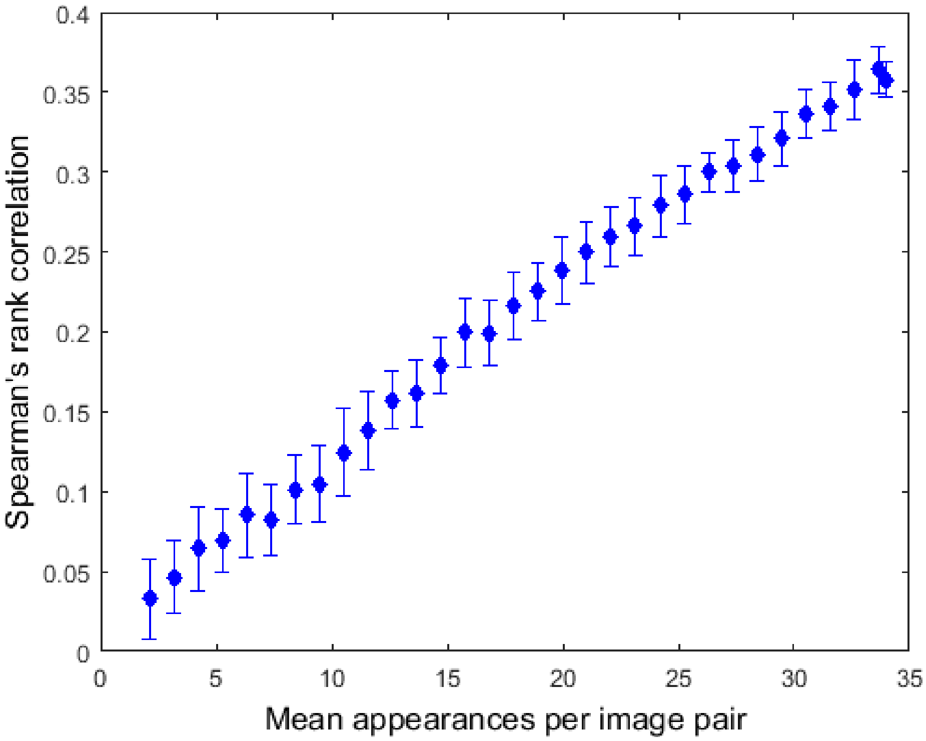


Figure F-5: *badlands* confusion rate consistency

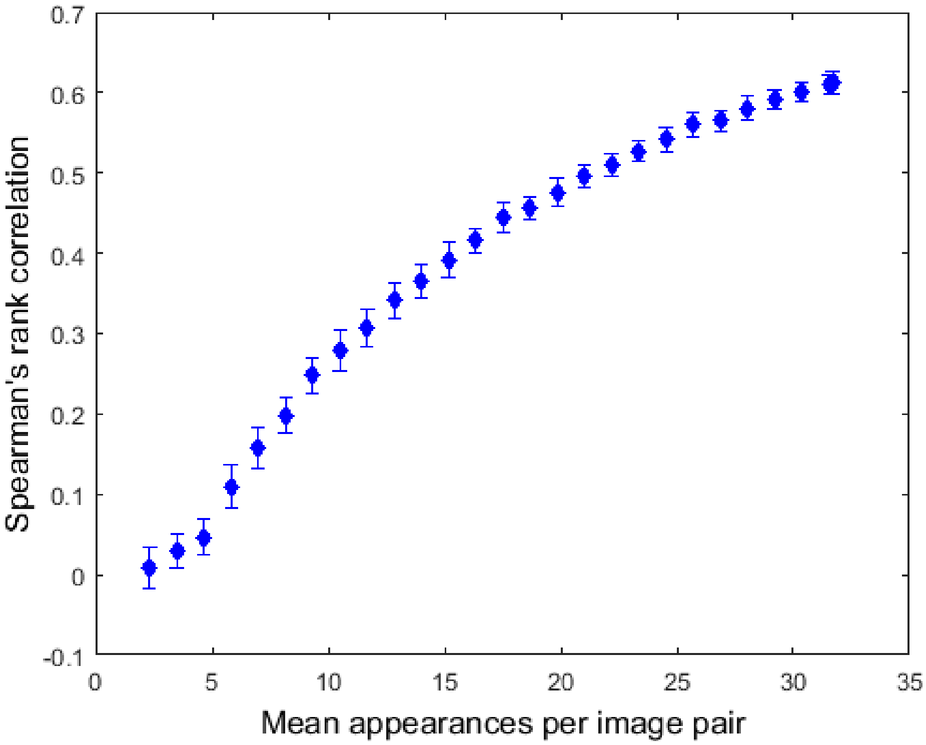


Figure F-6: *badlands* similarity score consistency

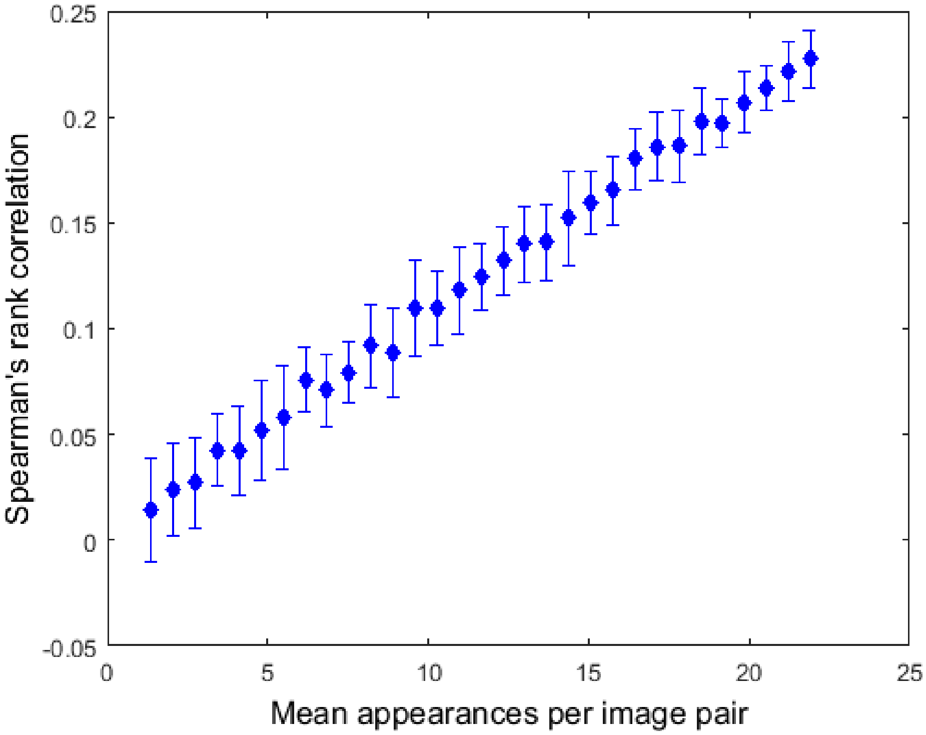


Figure F-7: *cockpit* confusion rate consistency

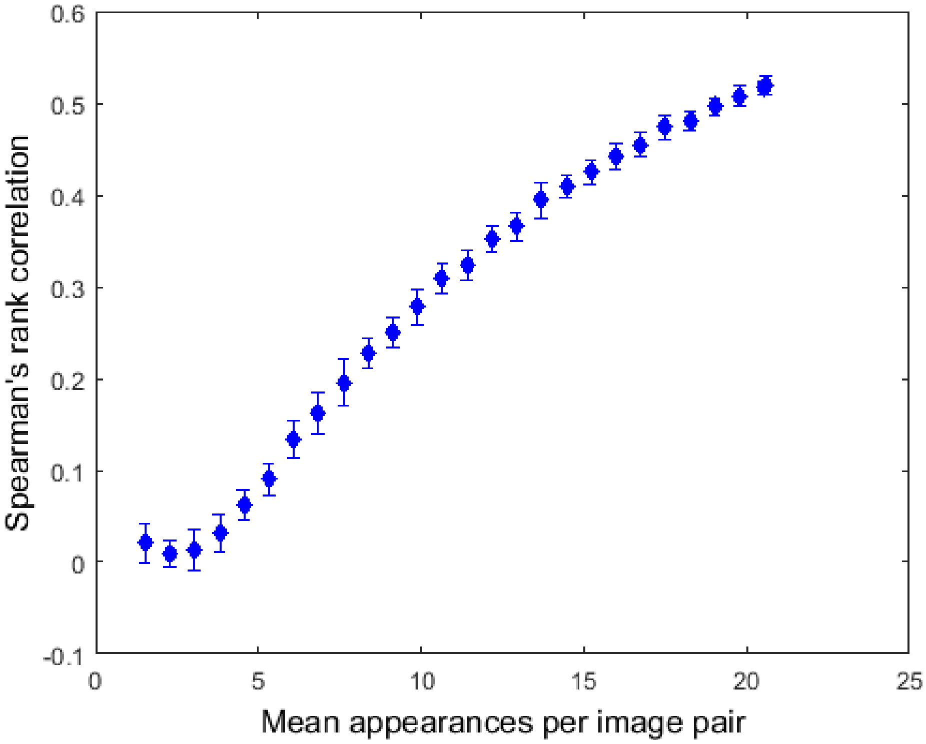


Figure F-8: *cockpit* similarity score consistency

Appendix G

Scatterplots for Neural Net Predictions

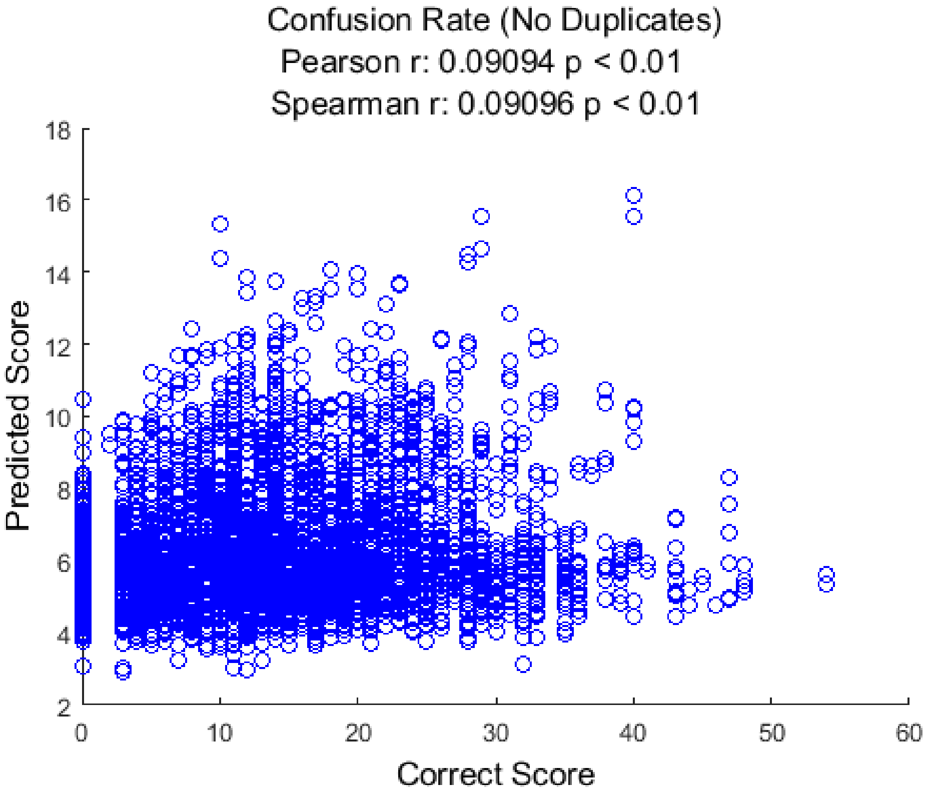


Figure G-1: Neural net prediction scatterplot for confusion rate

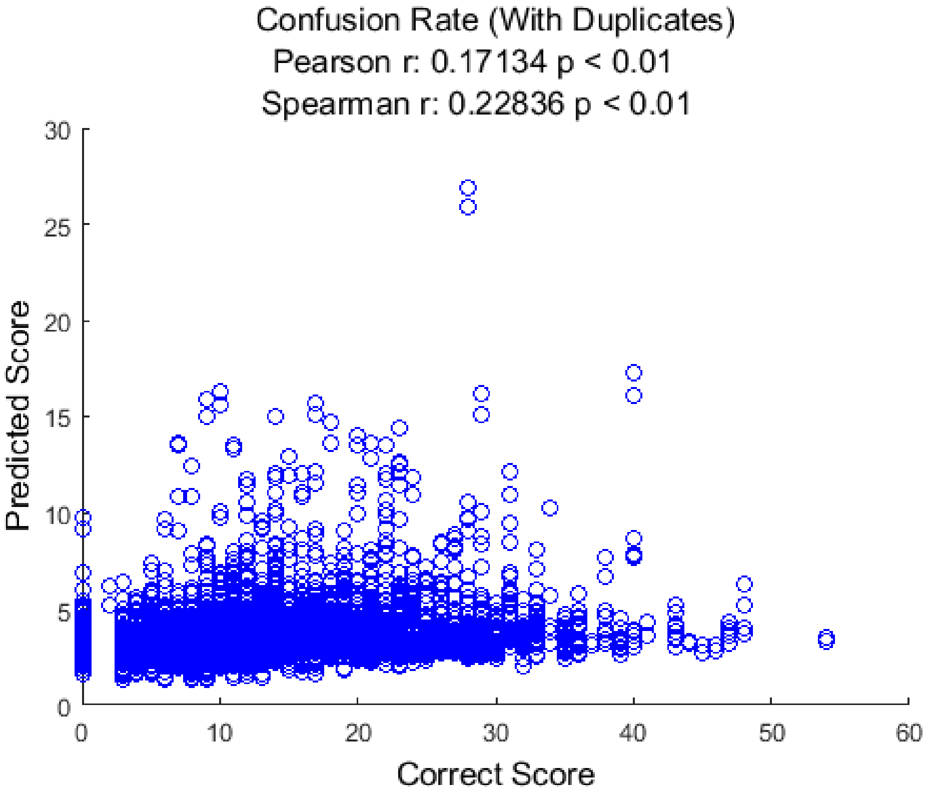


Figure G-2: Neural net prediction scatterplot for confusion rate with duplicates

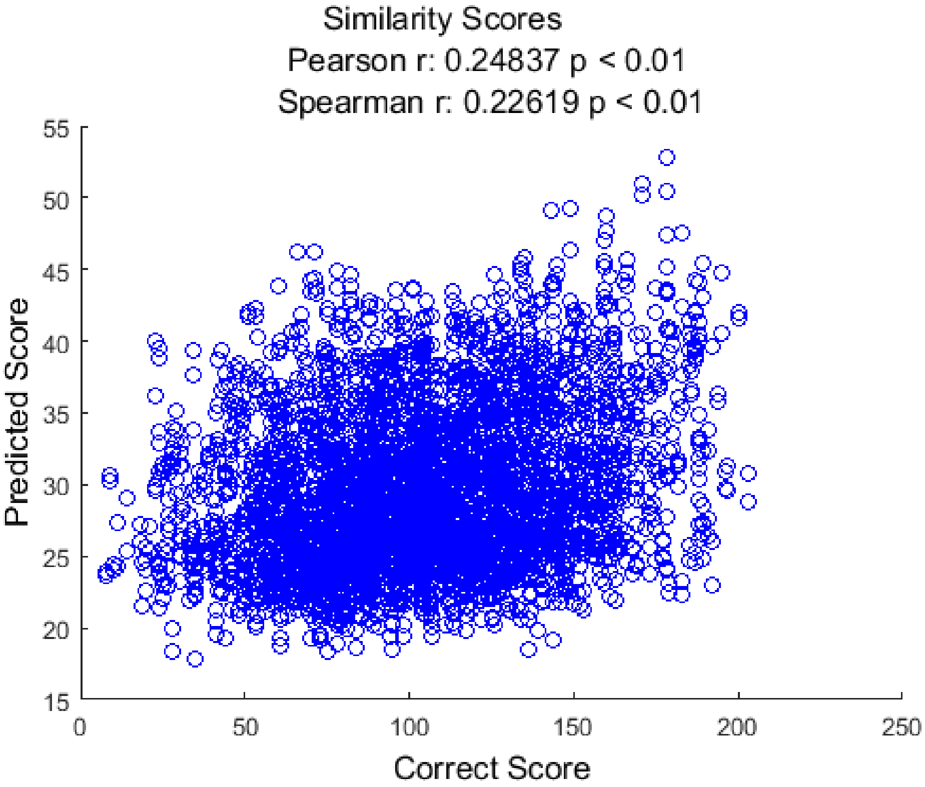


Figure G-3: Neural net prediction scatterplot for similarity score

Appendix H

Scatterplots for Neural Net Consistency

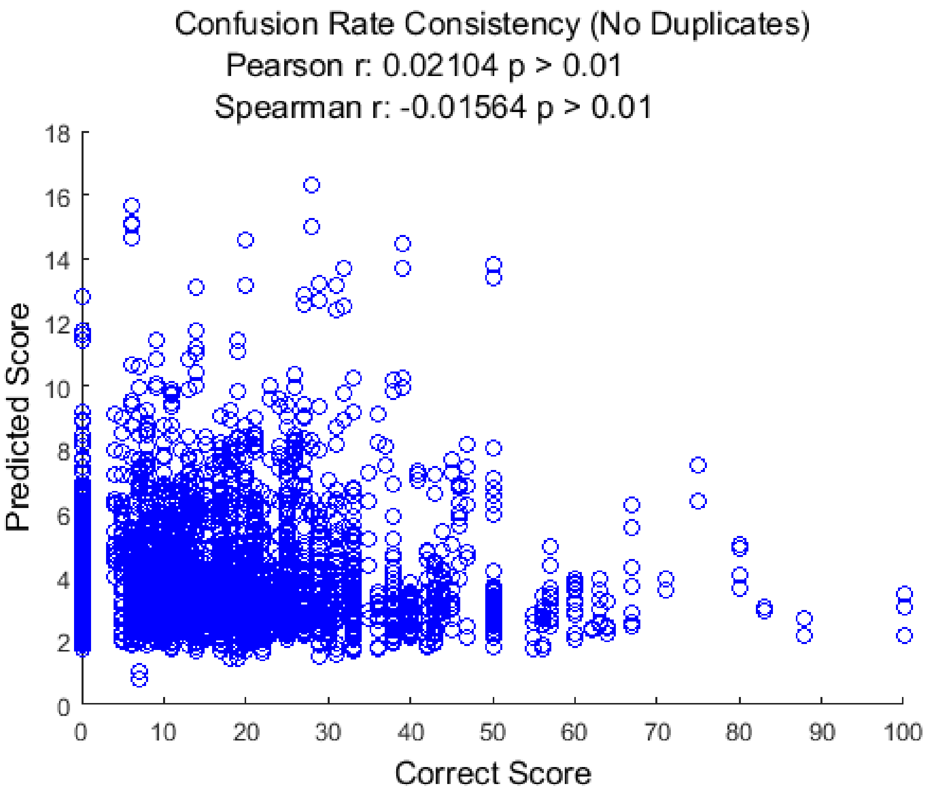


Figure H-1: Consistency neural net scatterplot for confusion rate

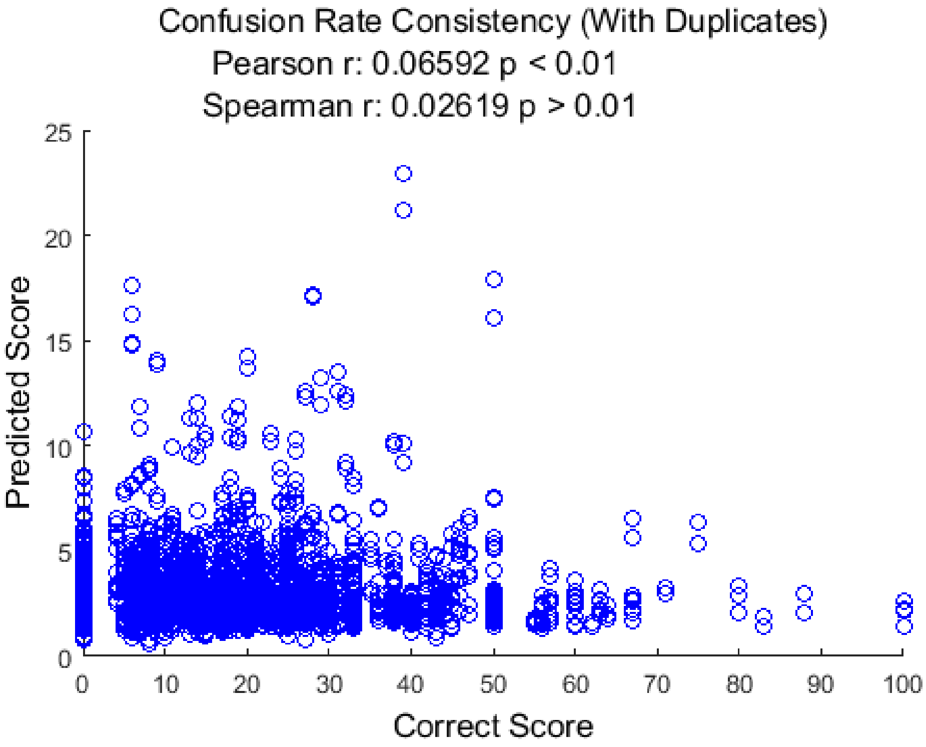


Figure H-2: Consistency neural net scatterplot for confusion rate with duplicates

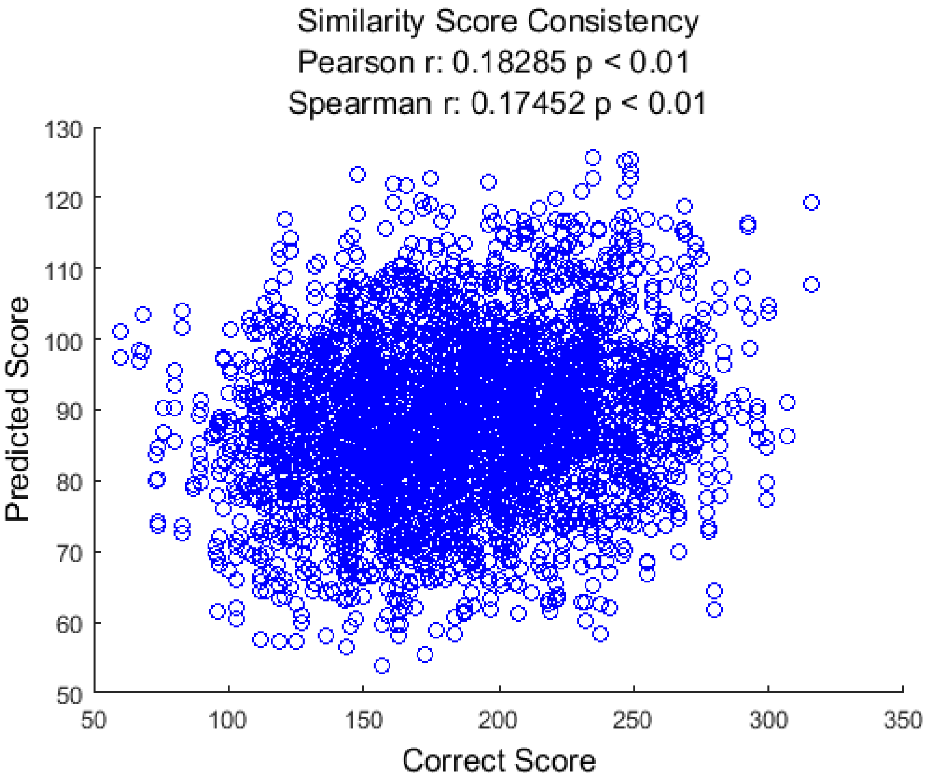


Figure H-3: Consistency neural net scatterplot for similarity score

Bibliography

[1] Cs231n convolutional neural networks for visual recognition. `http://cs231n.github.io/convolutional-networks/`. Accessed: 2015-5-22.

[2] W. Bainbridge, P. Isola, and A. Oliva. The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, 142(4):1323–1334, 2013.

[3] M. A. Borkin, A. A. Vo, Z. Bylinskii, P. Isola, S. Sunkavalli, A. Oliva, and H. Pfister. What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2306–2315, 2013.

[4] T. Brady, T. Konkle, G. A. Alvarez, and A. Oliva. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, 105(38):14325–14329, 2008.

[5] T. Brady, T. Konkle, J. Gill, A. Oliva, and G. Alvarez. Visual long-term memory has the same limit on fidelity as visual working memory. *Psychological Science*, 24(6):981–990, 2013.

[6] Z. Bylinskii, P. Isola, C. Bainbridge, A. Torralba, and A. Oliva. Intrinsic and extrinsic effects on image memorability. *Vision Research, special issue on Computational Models of Visual Attention*, 2015.

[7] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[8] E.Hoffer and N. Ailon. Deep metric learning using triplet network. *International Conference on Learning Representations*, 2015.

[9] M. Glanzer and J. Adams. The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16:5–16, 2010.

[10] A. Hirani, K. Kalyanaraman, and S. Watts. Least squares ranking on graphs. 2011.

[11] R. Hunter and J. Worthen, editors. *Distinctiveness and Memory*. Oxford University Press, 2006.

[12] P. Isola, J. Xiao, D. Parikh, and A. Oliva. What makes a photograph memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1469–1482, 2014.

[13] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? *IEEE Conference on Computer Vision and Pattern Recognition*, pages 145–152, 2011.

[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[15] A. Kholsa, A. Raju, A. Torralba, and A. Oliva. Understanding and predicting image memorability at a large scale. *Under Review.*

[16] M. Kibbe and E. Kowler. Visual search for category sets: Tradeoffs between exploration and memory. *Journal of Vision*, 11(3), 2011.

[17] T. Konkle, T. Brady, G. Alvarez, and A. Oliva. Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, 139(3):558–578, 2010.

[18] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012.

[19] Y. Lecun, L. Jackel, L. Bottou, A. Brunot, C. Cortes, J. Denker, H. Drucker, I. Guyon, U. Muller, E. Sackinger, P. Simard, and V. Vapnik. Comparison of learning algorithms for handwritten digit recognition. *International Conference on Artificial Neural Networks*, 1995.

[20] G. Ramanarayanan, J. Ferwerda, B. Walter, and K. Bala. Visual equivalence: Towards a new standard for image fidelity. *ACM Special Interest Group on Graphics and Interactive Techniques*, 2007.

[21] A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 512–519, 2014.

[22] F. Schmidt and J. Hunter. Theory testing and measurement error. *Intelligence*, 27:183–198, 1999.

[23] S. Schmidt. Encoding and retrieval processes in the memory for conceptually distinctive events. *Journal of Experimental Psychology: Learning, Memory, Cognition*, 11(3):565–578, 1985.

[24] P. Sinha and R. Russell. A perceptually based comparison of image similarity metrics. *Perception*, 40:1269–1281, 2011.

[25] R. Ulrich and J. Miller. Threshold estimation in two-alternative forced-choice (2afc) tasks: The spearman-kärber method. *Perception and Psychophysics*, 66(3):517–533, 2004.

[26] J. Vokey and J. Read. Familiarity, memorability, and the effect of typicality on the recognition of faces. *Memory Cognition*, 20(3):291–302, 1992.

[27] X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. 2015.

[28] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. *IEEE Conference on Computer Vision and Pattern Recognition, 2010*, 2010.