Automatic Triage for a Photo Series

Huiwen Chang[1] Fisher Yu[1] Jue Wang[2] Douglas Ashley[1] Adam Finkelstein[1]
[1]Princeton University [2]Adobe Research

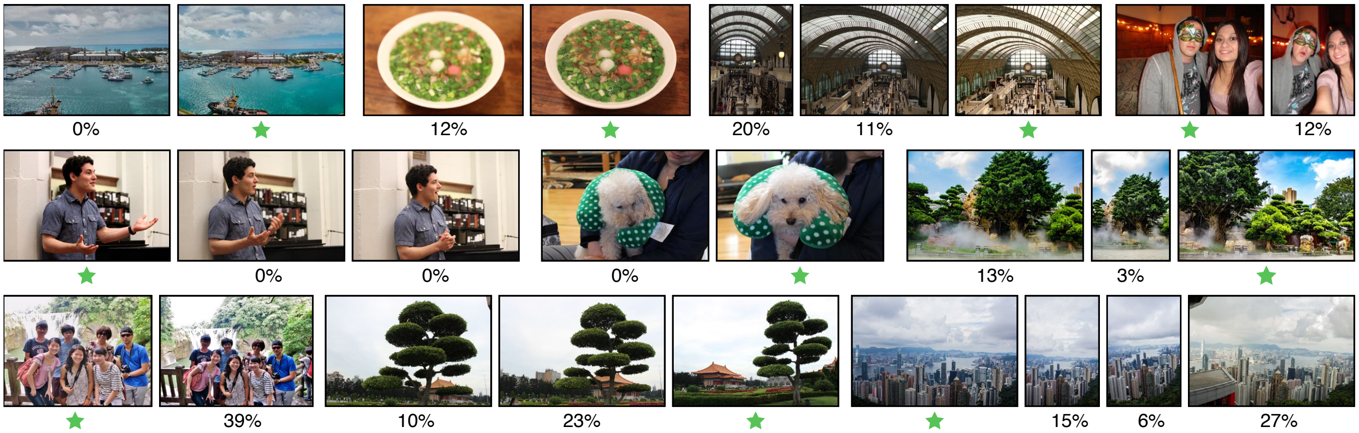


***Figure 1:** Multiple photo series, each sampled from one category in our dataset (Figure 3a: water, object, interior, selfie, ...). The starred photo in each series is the one preferred by the majority of people, while the percentage below each other photo indicates what fraction of people would prefer that photo over the starred one in the same series.*

Abstract

People often take a series of nearly redundant pictures to capture a moment or scene. However, selecting photos to keep or share from a large collection is a painful chore. To address this problem, we seek a relative quality measure within a series of photos taken of the same scene, which can be used for automatic photo triage. Towards this end, we gather a large dataset comprised of photo series distilled from personal photo albums. The dataset contains $15,545$ unedited photos organized in $5,953$ series. By augmenting this dataset with ground truth human preferences among photos within each series, we establish a benchmark for measuring the effectiveness of algorithmic models of how people select photos. We introduce several new approaches for modeling human preference based on machine learning. We also describe applications for the dataset and predictor, including a smart album viewer, automatic photo enhancement, and providing overviews of video clips.

Keywords: photo triage, photo quality, benchmark

Concepts: •**Computing methodologies** → **Computational photography;** *Neural networks; Image processing;* Ranking;

1 Introduction

The ease and ubiquity of digital cameras has led to an ever-increasing size of personal photo collections. To capture a vacation, party or other event, people often take a series of shots of a particular scene with varied camera parameters and content arrangement, in the hope of later being able to pick a few good photos to edit, post or share (Figure 1). However, sifting through and managing a huge collection of photos – grouping similar photos and deciding which are the "keepers" and which ones to omit – is cumbersome and time consuming. There are currently several commercial tools that facilitate this photo *triage* process. Photo browsers such as iPhoto and Picasa allow consumers to hierarchically organize and navigate through photo groups based on time or geographic information, while several online photo-sharing social networks like Facebook and Google Photos extract faces and higher level content for labeling purposes. While these tools improve album organization, none of them provide a way to automatically filter out the bad images or find the best images among a series of similar shots. To know which ones are worth keeping or sharing, users still need to sift through the entire album by hand, which is tedious. Our goal is to facilitate this process of finding the best images, by providing an automatic estimator of their relative ranking.

There has been extensive effort on assessing photo quality or aesthetics that gauges each image independently on an absolute scale (from "bad" to "good"). In contrast our goal is to establish a relative ranking for "better" or "worse" photos from among a series of similar shots (Figure 2). Existing methods that produce absolute scores tend to yield similar scores for similar images, and therefore perform poorly for our problem. There are also methods

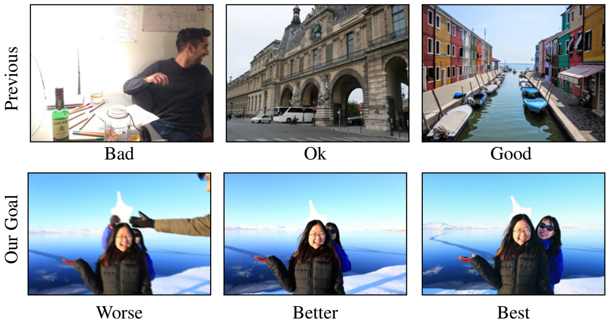


***Figure 2:** Above: Previous work evaluates photo quality on an absolute scale (bad to good). Below: Our goal is to find relative ranking among a series shots of the same scene (for which previous methods would typically provide similar absolute scores).*

SIGGRAPH '16 Technical Paper,, July 24 - 28, 2016, Anaheim, CA,
ISBN: 978-1-4503-4279-7/16/07
DOI: http://dx.doi.org/10.1145/2897824.2925908

for selecting the best among a photo "burst" (typically based only on low-level features like blurriness or closed eyes) but in our case the composition usually varies substantially more and requires higher-level features for effective comparison.

To address this problem, we collected a new dataset that contains $15,545$ personal photos, organized in $5,953$ series. The dataset consists of unedited photos from personal photo collections. The contributors are photographers of varying ability, and the vast majority of the photos are "casual" – characteristic of typical consumer photo albums. We conducted a study on Amazon Mechanical Turk in which subjects were asked for preferences (and comments) among $98,780$ pairs of photos, each pair from within a series, from which we extract a ground truth ranking of human preference across each series (★ and % in Figure 1). With this data we also establish a benchmark – training, validation and test sets, as well as a set of criteria for evaluating new models for human preference against the data. We believe this is the first large public dataset of unedited personal photos designed to address the photo triage problem.

To model human preference, we experimented with several machine learning approaches based on hand-crafted features as well as features established in the object recognition literature. Among pairs of images where human preference is reasonably consistent ($\geq$70% agreement), our best-performing model (based on a variant of the Siamese network prototyping from the VGG model) is able to predict human preference 73% of the time. Comparing with several baseline methods from previous research, we find the best-performing one (from Khosla et al. [2015]) achieves 57% accuracy.

In addition to modeling human preference for automatic triage, we also introduce two new applications for this kind of data. First, we show that new images can be automatically enhanced by analogy to other pairs of images in the dataset. Second we show that our learner can be used to select "good" frames from video to provide an overview of a shot.

2 Related Work

Photo Triage. With the ever-increasing size of personal photo collections, photo triage has drawn the attention of both researchers and developers in recent years. Drucker et al. [2003] design a user interface which enhances viewing experience and eases the control for users during photo triage. Jacobs et al. [2010] introduce the notion of "cosaliency" – local structural changes between image pairs – which facilitates the manual triage process by helping the viewer easily see the differences between images. Both of them aim at expediting the process of photo triage interactively, while we propose a fully automatic approach for triage among photo series. Zhu et al. [2014] focus on selecting more attractive portraits from large photo collections based on facial features. In addition to considering facial expressions, our paper addresses the general (more difficult) problem, including natural scenes, urban environments, interior scenes, cluttered scenes, and lone objects.

Image Summarization. Researchers have worked on summarizing image collections The primary goal of this line of research is to select a few representative images from a collection by jointly considering photo quality, event coverage and scene diversity. Sinha et al. [2011] propose content- and context-based optimization functions for summarizing large personal photo collections. Simon et al. [2007] propose a summarization approach that is more tailored towards 3D scenes. The triage problem we address can be viewed as a sub-problem of image summarization, as it focuses only on finding good images from photo series, but does not consider the representativeness of these images to the whole album, an essential consideration in summarization. Our approach thus can be used as a component to improve existing summarization systems.

Quality Assessment. An alternative way to deal with photo triage is selecting by assessing quality for each photo in series. Researchers have studied aesthetics or quality evaluation for photos, and such methods can be used to determine a ranking among photos. Early efforts employ handcrafted features to evaluate visual aesthetics and adhere to principles of photography. The features involve both low-level features such as lighting [Luo and Tang 2008; Bychkovsky et al. 2011; Kaufman et al. 2012; Yuan and Sun 2012], texture [Datta et al. 2006; Tang et al. 2011] and color [Datta et al. 2006; Nishiyama et al. 2011], as well as high-level features such as composition [Luo and Tang 2008; Bhattacharya et al. 2010; Liu et al. 2010; Dhar et al. 2011; Guo et al. 2012; Park et al. 2012; Zhang et al. 2013] and content [Dhar et al. 2011; Luo et al. 2011; Kaufman et al. 2012]. Such approaches achieve good results by using joint features to predict image aesthetics, but they are typically limited to heuristics inspired by the photographic guidelines (e.g., the rule of thirds, golden ratio, and simplicity). Because generic visual features are effective to represent semantic image descriptors, methods like that of Marchesotti et al. [2011] and Murray et al. [2012] attempt to rate visual properties using features such as SIFT [Lowe 2004], GIST [Oliva and Torralba 2001], and the Fisher Vector [Marchesotti et al. 2011]. Ye et al. [2012] borrows the idea of code book representations and automatically predicts human perceived image quality without a reference image. Recently, Lu et al. [2014] applied deep neural network approaches and outperformed handcrafted and generic visual features on an aesthetic dataset. For the benchmark introduced in this paper, the best existing method we have found is the "memorability" score of Khosla et al. [2015], which is out-performed for this task by new approaches we introduce. The general strategy of establishing an absolute assessment across the range of images found in a typical photo album is more difficult than finding a relative ranking among a series of similar photos (Figure 2), and existing methods for finding absolute scores tend to provide similar scores across a series.

Datasets for Aesthetics Assessment. Prior datasets for aesthetic analysis of photos have been obtained from online photo sharing communities. Photography enthusiasts, including both amateur and professional photographers, share photos and rate those taken by peers in these communities. Datta et al. [2006] compiled a dataset containing more than 3,000 images from Photo.net, with scores ranging between 1 (ugly) and 7 (beautiful). The CUHK dataset of Ke et al. [2006] contains 60,000 images from DPChallenge.com, with binary labels (good or bad). The AVA dataset of Murray et al. [2012] compiles aesthetic judgments from on-line communities of photography amateurs to obtain 250,000 images with human rated aesthetic scores. It has been used for evaluation in the most recent aesthetic assessment research. The dataset presented in this paper complements these prior efforts in three ways. First, our data are collected from raw photo albums (rather than harvested from online sharing communities) so the photos themselves are characteristic of personal albums – they have not been prefiltered for quality by the owners, nor have the photos themselves been edited. Second, our photos are organized into a collection of photo series. Third, our images are complemented by crowdsourced human judgements that allow us to rank photos within a series. These three attributes uniquely tailor our dataset to the problem of triage among photo series.

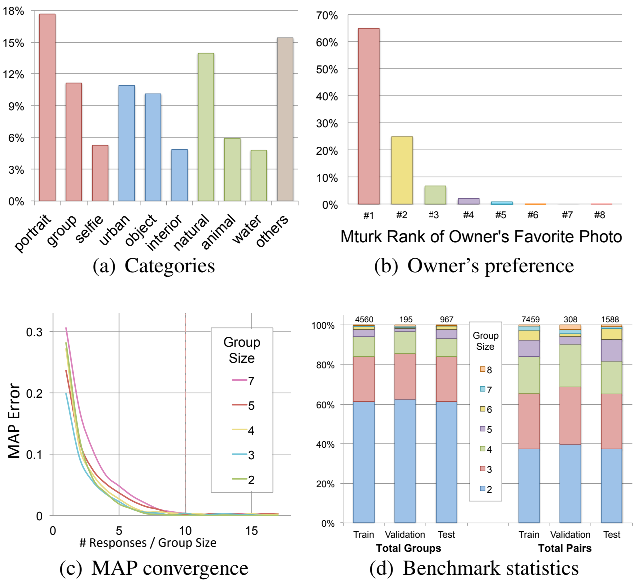
3 The Dataset

To understand how human beings evaluate similar photos, we first collect $5,953$ photo series from a large group of contributors, containing $15,545$ personal photos. We then conduct a crowdsourced user study on people's responses in these series that contains $98,780$ preferences and rationales in total. In this section, we first describe the methods and apparatuses of our data collection and user study process in more detail, then provide an analysis of the user study results.

3.1 Collecting Personal Photo Series

Since there's no public dataset to address the problem of photo triage for a series of shots, we describe in this section how we build a dataset specially targeted at this problem. But there are strong restrictions concerning the availability of data imposed by our goals. For example, we intend to collect unedited, complete photo series from personal albums. Such data is hard to harvest online, since most online public photo sharing sites (e.g. Flickr) only contain images that the users have selected to submit, and most users only select a small number of good photos from their photo series to share. Furthermore, a large portion of images submitted to these sites have already been edited or enhanced. Other cloud storage services such as Apple iCloud do store users' raw photo series, but their data is inaccessible. In order to collect a large scale personal photo series dataset for open academic research, we designed and ran a contest that was open to college students, faculty, and staff. In the contest, we asked participants to submit unedited or slightly edited personal photo albums, where image sizes are larger than 600 × 800 pixels. The contest lasted for two weeks, and we have collected over 350 album submissions from 96 contributors.

To identify photo series from all the submitted images, we first discard redundant shots, e.g. those captured in a burst session, as the differences among them are too subtle to be treated as a series. We achieve this by sorting images according to the dates in their metadata, and computing the Root Mean Square (RMS) color difference between neighboring images downsampled to 128^2. Duplicates are identified when the RMS difference is smaller than a threshold (0.06 for normalized colors in $[0, 1]$, found by experimentation).



(a) Categories

(b) Owner's preference

(c) MAP convergence

(d) Benchmark statistics

Figure 3: *Dataset properties. (a) Distribution of categories is consistent with personal photo collections. (b) Turkers generally agree with the photo contributors' favorite photo in a series. (c) MAP error declines with increasing number of human comparisons, converging by ten times the group size. (d) Distribution of group sizes and labeled pairs within the training, validation and testing sets of our benchmark.*

To find photo series from remaining images, we extract SIFT descriptors and build scene correspondences based on SIFT matching [Lowe 2004] between neighboring images. If two neighboring images have good scene matching, we group them together into a series. However, local feature matching cannot handle special cases such as one of the two images having severe camera motion blur. To handle such cases, we also check the color similarity between two images as another metric to merge images into series, measured by the earth moving distance of the color histograms. Finally, since a large portion of user photos contain human faces, we apply face verification using the Face++ Toolkit [Megvii Inc. 2013; Zhou et al. 2013] in all the merged photo series. For semantic consistency, we expect each series contain the same group of people. If a series contain different groups in different images, we split it into smaller ones based on face identity. We also do not allow a series to have more than 8 images. If a series contains more than 8, we split it by using variant k-means on the 116-dimension global features as depicted in [Wang et al. 2013], where the center is a representative photo instead of a mean. After gathering all the clusters automatically using the process described above, we manually checked the resulting series to filter clusters with terrible image quality or privacy concerns (such as credit cards).

This data collection and selection process yields $5,953$ photo series: a few examples are shown in Figure 1. Together with Figure 3(a), they show that our dataset contains a wide variety of photography subjects, including portraits, group and family photos, selfies, urban scenes, objects and food, interiors, natural landscapes, animals and water scenery. They also show that images in our dataset contain large variations in composition, human pose, color and lighting characteristics, etc.

3.2 Crowdsourced Ranking

To obtain human preference on images in each photo series, we employ the Amazon Mechanical Turk (MTurk) [Kittur et al. 2008] for a crowd-sourced user study. Given that our goal is to develop a relative photo quality metric instead of an absolute one, we ask participants to perform pairwise comparisons on images from the same series. For each comparison, a pair of photos are randomly selected from a series and are shown side-by-side, each fitted into a 640x640 frame with the original aspect ratio. The question we ask for each pair is: "Imagine you take these two photos, and can only keep one. Which one will you choose, and why?" We used a forced-choice methodology in order to better measure small differences. Participants are also required to fill out at least one of the two forms: one describes the reason(s) why a particular photo is preferred (positive attributes); the other describes the reason(s) why the other photo is not preferred (negative attributes). The purpose of asking for written explanation is two-fold: (1) it forces the participants to make justifiable decisions, thus the quality of gathered user data is higher; (2) the user comments carry insightful information that can guide us on feature extraction.

The pairwise comparison results are used to obtain a global ranking for each image in a series. Denote the pairwise comparison annotations as a count matrix $S = \{s_{i,j}\}$, where $s_{i,j}$ is the number of times that the ith photo I_i is preferred over the jth photo I_j. We use the Bradley-Terry model [1952] for obtaining the global ranking, which describes the probability of choosing I_i over I_j as a sigmoid function of the score difference between two photos $\Delta_{i,j} = c_i - c_j$, i.e.

$$P(I_i > I_j) = F(\Delta_{i,j}) = \frac{e^{\Delta_{i,j}}}{1 + e^{\Delta_{i,j}}} \qquad (1)$$

The score parameter c can be estimated by solving a maximum a-posteriori (MAP) problem. Since we assume the prior is a uniform distribution, the objective is to maximize

$$\log Pr(S|c) = \sum_{i,j} s_{i,j} F(\Delta_{i,j}) \qquad (2)$$

which could be solved by using gradient descent.

Previous work [Zhu et al. 2014] shows that the convergence of the MAP estimation typically occurs in a linear rather than a quadratic number of pairwise comparisons. We thus conduct a pilot user study to determine the linear coefficient k, i.e. $k(n-1)$ pairwise comparisons need to be collected when the size of the series is n. In the pilot user study, we evaluated the MAP estimation on 15 photo series of different sizes with varying numbers for the linear coefficient. As illustrated in Figure 3(c), all MAP averages in different series converge by $k = 10$. Thus in our subsequent study, we gather $\lceil 20/n \rceil$ responses per pair in a series of n photos.

3.3 Analysis of Crowdsource Data

We have received $98,780$ responses from MTurk in total. We briefly report some interesting findings from the results here.

Preferences

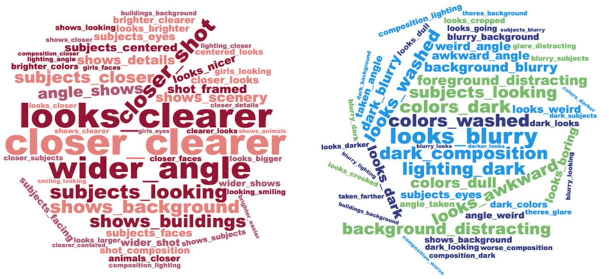
It is clear that people have different levels of consistency on evaluating different image pairs. In Figure 7, we show example pairs of different levels of agreement. For pairs that have clear human preferences, the compared images often differ significantly in some image features, making automatic prediction easier to succeed. For examples with more diverse options, it is a harder job for an automatic algorithm to make a decision that is consistent with the majority. Therefore, the problem of automatically predicting human preference becomes easier as the level of agreement increases (Figure 7b-f). Difficulties are well balanced across our dataset (top of bars).

One may argue that evaluating personal photos can be subjective, depending on whether or not the viewer has prior knowledge about the scene/people in the photos. We conducted an additional experiment to investigate whether MTurk reviewers' preferences are consistent with those of the photograph owners. We selected 551 sampled series from 10 owners, and asked the album owners to select the ones they like most from their own photo series. The results are shown in Figure 3(b). In 358 out of 551 series, the owners made exactly the same decisions with the collected intelligence of MTurk reviewers. In addition, in 69.5% of the series in which album owners made different choices, the MTurk reviewers also had relatively low consistency among themselves. We thus conclude that MTurk responses offer a reasonable proxy not just for generic human preferences but also for photo owners, on our dataset.

Comments

To better understand the determining factors when people compare similar images, we extract the most frequent double-word phrases from the comments, including both positive and negative one, as visualized in Figure 4.

For negative reasons, blur, dark lighting, washed colors, and distracting foregrounds or backgrounds were the primary topics of discussion. For positive reasons, less blurry, clearer and closer photos showing more detail, and brighter colors were highlighted. People also liked wide angle shots often. These findings provide a useful guidance for defining features for an automatic recommendation system.



(a) Reasons for preferred photos (b) Reasons for rejected photos

Figure 4: Word clouds visualizing most frequent rationales for photo preference.

3.4 Benchmark

To facilitate future research, we set up an online benchmark for photo triage which may be found together with our dataset and models at our project webpage phototriage.cs.princeton.edu. We divide the whole dataset into three subsets. Out of the 5,953 series, 4560 are randomly sampled for training, 195 for validation, and the remaining 967 for testing. Figure 3(d) shows a detailed breakdown for each set. Upon the acceptance of the paper, we will release all of the images but only the human labels for the training and validation sets will be publicly available. New results on the testing set can be submitted through an online interface, and will be evaluated by two metrics: log likelihood and accuracy, which will be explained in more detail in Section 5.

4 Modeling Preferences

Based on our dataset, we propose a variety of methods for learning and predicting human preference in evaluating photo series, including both feature-based approaches (Section 4.1) and end-to-end deep learning (Section 4.2). In Section 5, we quantitatively compare these methods with previous approaches in the literature.

4.1 Feature-based Learning

Extensive work has been done on extracting or learning features for photo aesthetics assessment [Nishiyama et al. 2011; Kaufman et al. 2012; Zhang et al. 2013]. To address the new problem of learning human preference between a pair of similar images, we first explore assorted features motivated from either heuristics or prior work, including color and lighting statistics, clarity, face factor, composition, and content.

Hand-tuned Features

Color and Lighting As revealed by our user study results, people are usually sensitive to color and lighting differences between two photos of the same scene. For measuring them we compute a standard color histogram feature of 16 bins in L*, a*, and b* channels in CIELAB color space.

Face. From the MTurkers' comments on portrait photos, the dominant facial features are "smiling face", "closer face", "face angle", and "eyes". We detect face position, areas, angle, smileness and eye openness using existing methods [Cootes et al. 1998; Cao et al. 2014], and normalize them to $[0, 1]$. Since multiple faces could be detected, we extract the median of all face positions, both the median and max of face areas, angle, smileness, and eye openness as the face feature.

Composition. In previous work, general composition rules such as the rule of third and the rule of diagonals are widely used for photo aesthetics evaluation [Liu et al. 2010; Dhar et al. 2011]. These approaches detect the salient objects as foreground regions in an image, and use their locations for evaluating composition. In our problem, aside from detecting salient regions in a single image, we need to further discover the common salient region that appears in both images, and explore its composition difference. To achieve this, we first segment both images into 600 superpixels using Ren et al.'s method [2003]. We then compute pixelwise saliency maps using the method proposed by Judd et al. [2009], and compute the average saliency for each superpixel. Next, we build non-rigid dense correspondence between two photos [HaCohen et al. 2011], and select the common foreground region as a set of salient superpixels that appear in both images. We encode the foreground saliency map as a 900 dimension vector by downsampling to 30×30, and use it as the composition feature.

Clarity. Having motion blur or being out of focus is a common problem in casual photographs. For blur detection, we compute the clarity score using the CORNIA metric [Ye et al. 2012], in both the entire image and the extracted shared foreground region. This is motivated by the observation that a sharp foreground against a blurry background is often visually pleasing, although its overall clarity score could be low.

Deep Features

Content. Convolutional networks (ConvNet) have been shown to produce high quality image representations. Deep features trained for object classification have been successfully applied to a variety of content-related computer vision tasks such as object recognition, object detection [Girshick et al. 2014; Ren et al. 2015b], and recognizing image styles [Karayev et al. 2013]. We take features pre-trained on the ImageNet dataset via AlexNet [Krizhevsky et al. 2012] and the 16-Layer VGGNet [Simonyan and Zisserman 2014], and extract the response of the second-to-last fully connected layer as our content features. Both feature sets have 4096 dimensions and are computed from center crops of images resized to 256^2.

Combining Features

We concatenate the hand-crafted features with deep features described above from a given photo pair, and preprocess the feature vectors by a whitening transformation. We then train a Random Forest classifier [Breiman 2001] and Support Vector Machines (SVM) with Ridge normalization to learn which image in the pair is more preferable as a binary label. In the Ridge SVM experiment, we set alpha to 1.0 and use the LSQR solver [1982]. In the Random Forest method, we fit 100 trees on the sub-samples of the dataset. We evaluate the performance of hand-tuned features, deep features, and the fusion of both in the Section 5.

4.2 End-to-End Learning

In this section, we explore using ConvNet to design an end-to-end prediction model. Recent success on image representation learning illustrates the power of learning image features directly. Since AlexNet was used in ImageNet image classification challenge, convolutional neuron networks have dominated the challenge leaderboard and been constantly improved [Szegedy et al. 2014; Simonyan and Zisserman 2014; Szegedy et al. 2015; He et al. 2015]. The main benefit of using convolutional neuron networks is that they can learn the image representation and prediction model directly (end-to-end learning). The success of convolution networks demonstrate that the learned representation can outperform the hand crafted features in the classification task.

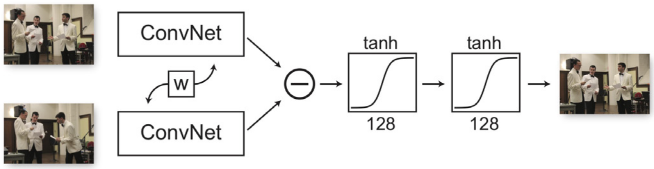


Figure 5: Siamese architecture: Features are first extracted from each of the pair of images by two ConvNets with shared weights. Then the difference of the features are passed through two levels of hidden layers, with 128 channels activated by `tanh`*. Finally a two-way* `softmax` *decides which image is preferred.*

A natural question is how we can learn image representation for our task. The main difficulty is lack of data. Although we have more than 10,000 images in our dataset, it is still far less than the number of images used to successfully train a deep convolutional network from scratch. Because the network for learning natural image representation usually requires millions of parameters, network training without large amount of data can easily get stuck in bad local minimum. This problem also exists for most of the computer vision tasks. While there are millions of images available for classification, there are fewer images for the other tasks such as object detection and segmentation, since it is much harder to obtain accurate ground truth annotations. One way to solve the training problem is to have good initialization. To achieve this, researchers have tried to adopt the network designed for image classification together with the model trained on the millions of images to new tasks such as object detection [Girshick 2015; Ren et al. 2015a], semantic segmentation [Long et al. 2014; Yu and Koltun 2015], memorability [Khosla et al. 2015] and so on. The adoption involves two steps. The network is first extended to a new domain by adding new components [Long et al. 2014; Hariharan et al. 2014; Yu et al. 2015]. For example, for semantic segmentation in [Long et al. 2014], instead of predicting a single label for the whole image, the classification network can be extended to predict a label for each pixel by adding upsampling layers. Then in training, the extended network is initialized by the model parameters pretrained on the image classification task, and then it is fine-tuned with the data for the new task. This fine-tuning scheme is working very well and the resulting algorithms can achieve better performance than those using hand-crafted features. Furthermore, it is found [Yu and Koltun 2015] that even if the original network is slightly modified, the pretrained model parameters can still act as effective initialization. Therefore, we try to design a network based on existing architecture and fine-tune the network parameters on our dataset.

To extend the image classification network, we have to consider several differences of our problem. Firstly, the networks for image classification only take one image as input and make the prediction, while our prediction depends on a pair of images. Also, the features for an image should be the same no matter whether the image is the first or second in the pair. This leads us to consider Siamese architecture to learn extracting image features. In Siamese architecture [Bromley et al. 1993], two inputs are sent into two identical sub-networks independently, which share the same network parameters in both training and prediction phases. As shown in [Bromley et al. 1993; Chopra et al. 2005; Bell and Bala 2015], Siamese network can learn image embedding when the supervision is the distance between two inputs. However, in our problem, we only have a binary label for which input is preferred. Also, if we swap the two inputs, the label is supposed to be flipped. This motivates us to design a new cost function based on Siamese architecture.

We now put forth the idea more concretely. The input of our model is a pair of images $(I_1, I_2) \in \mathbb{I} \times \mathbb{I}$, where $\mathbb{I}$ is the space of images. We aim at learning a function $p : \mathbb{I} \times \mathbb{I} \mapsto \{-1, 1\}$, where 1 means the first image is better and -1 means the opposite. The definition requires p be skew-symmetric, that is, $p(I_1, I_2) = -p(I_2, I_1)$.

To achieve this, we decompose the prediction function into two stages $s : \mathbb{I} \times \mathbb{I} \mapsto \mathbb{R}^n$ and $f : \mathbb{R}^n \mapsto \{-1, 1\}$ so that $p(I_1, I_2) = f(s(I_1, I_2))$. Conceptually, s is the n-dimension feature extraction function learned from the input image pairs and f classifies the features computed by s. We observe that if s is skew-symmetric and f is odd, p is skew-symmetric. We use a Siamese architecture to learn s, and use a multi-layer proceptron to learn f. The details of the two stages are explained in the following two paragraphs.

At the first stage of learning s, two input images are first passed to two identical sub-networks with shared weights (parameters). The final output of the first stage is the difference of the outputs of the two identical sub-networks with different inputs. Different from most of the other uses of the Siamese architecture, we don't compute the distance of the outputs from the identical networks, because our label for the input image pair doesn't tell us how similar or dissimilar they are. Instead, we pass the difference to the second stage to classify which image is better. In our implementation, we try both AlexNet and 16-layer VGGNet as the sub-networks and initialize them with the weights trained on ImageNet dataset. However, the last fully connected (FC-1000) and soft-max layers are removed. So the output dimension of each sub-network is 4096. To guarantee that the sub-network can produce the same results when they take the same inputs, Dropout layers are also removed. We find this is helpful for training the network.

At the second stage of learning f, a multi-layer perceptron classifies the features of the image pairs. In our perceptron, there are two hidden layers, each of which has a 128-dimension output. Each hidden layer comprises a linear fully connected layer and a non-linear activation layer. Because f is supposed to be odd, we use $\tanh$ in activation layers. The outputs of the second hidden layer are fed to a two-way Softmax. The outputs of Softmax indicate which of the two input images is better. We do not use Dropout for regularization because we find that Dropout can prevent the training from reaching convergence in our experiments.

We concatenate these two stages of network and train them together. AlexNet and VGGNet take input images of size 227^2 and 224^2, respectively. Therefore, we resize each image so that its larger dimension is the required size, while maintaining the original aspect ratio and padding with the mean pixel color in the training set. Stochastic gradient decent (SGD) is used to optimize the network with L2 regularization. The learning rate is 0.001. The momentum is 0.9 and weight decay is 0.0005.

5 Results

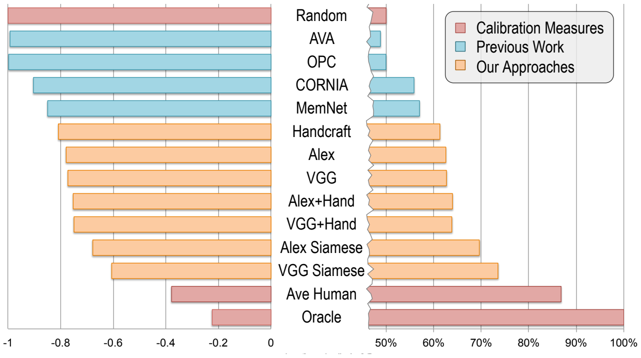
We evaluated different methods at two levels: series-level and pair-level, both visualized in the pair of bar charts in Figure 6. The series-level evaluation measures the log likelihood orderings over all series as follows. For a series k, given the human preferences on each image, we identify the one most preferred by humans as $winner(k)$. For a method M, we apply it to all image pairs that include $winner(k)$ and another image in that series. Next we compute the logarithm of the joint probability of the decisions of method M on all such pairs.

$$\log \mathcal{L}_k(M) = \sum_{i \neq winner(k)} \log P(M, winner(k), i),$$

where

$$P(M, i, j) = \begin{cases} Pr(I_i > I_j), M \text{ predicts photo i better than j} \\ Pr(I_j > I_i), \text{otherwise} \end{cases}$$

$Pr(I_i > I_j)$ is calculated according to Bradley-Terry model (see Section 3.2). The resulting log-likelihood value is negative, and visualized in Figure 6 relative to a naive baseline of random guessing. The reason we use the log-likelihood measurement instead of the frequency with which a particular method selects the best image is that series of different sizes have different levels of difficulty – it is harder to choose the best one from larger series than smaller ones.



(a) Log likelihood (b) Accuracy ($\geq$70% agreement)

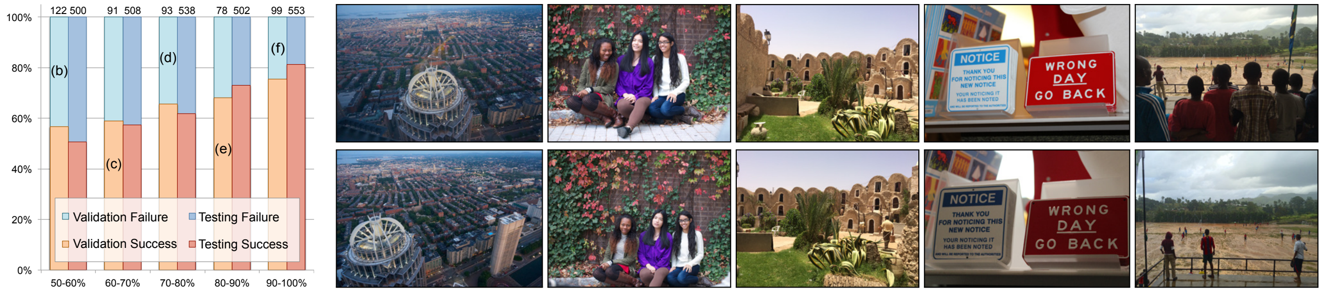
Figure 6: *Benchmark performance of various methods, where further to the right indicates better performance in all cases: (a) Log likelihood of predictions made by each method, based on human preferences in the testing set, where values are normalized such that random guessing has value -1. (b) Each method's accuracy in predicting human preference among pairs where humans agree at least 70% of the time. Overall, VGG Siamese is our best performing method, beating our other proposed approaches as well as baseline methods from previous work.*

On the pair level, considering that some pairs have a clear human preference while others do not (Figure 7), we only choose pairs where the majority agreements is over 70% calculated by Bradley-Terry model as test pairs, and compute the accuracy of different methods on these high-consistency pairs.

In both results shown in Figure 6, the "oracle" predictor selects the majority of human votes, while "average human" selects each photo with the probability proportional to the human votes, which represents the average performance of a single person.

We also identify several previous approaches for comparison. CORNIA [Ye et al. 2012] is a codebook-based approach for non-reference image quality assessment. This work deals with general low-level quality attributes such as noise and blur. Given a photo pair, CORNIA provides a distortion score for each photo and the one with the lower score is considered to be better. Although it uses only low-level features, it performs reasonably well on our dataset, given that image sharpness is a main factor for comparing similar images, according to the user study results.

RAPID [Lu et al. 2014] uses a deep neural network to evaluate the aesthetic quality of a given image. Taking cropped and resized photos as features, it produces a probability of whether a given image is high-quality. We use this probability to make the decision when comparing an image pair. While their approach works well on the AVA dataset [Murray et al. 2012], its performance in our task is close to random guess, which implies that their neural network is incapable of differentiating images that are in similar aesthetic levels and image styles. Liu et al. [2010] proposed OPC, which uses several aesthetic guidelines for evaluating photo composition. We use this score as a criteria to compare the aesthetics quality of an image pair. The performance of this method on our dataset is also close to random guess. This is because composition is only one factor among many others when people evaluate similar images, and a few photography rules are not comprehensive enough for describing the subtle composition differences among casual photos.



(a) degree of human agreement (b) $\sim$ 55% failure (c) $\sim$ 65% success (d) $\sim$ 75% failure (e) $\sim$ 85% success (f) $\sim$ 95% failure

Figure 7: *Photo pairs in validation and testing sets that have greater human agreement tend to be easier to predict. (a) The performance of our best predictor (Siamese) relative to increasing deciles of human agreement. (b-f) Examples of prediction failures and successes from these deciles, where the upper photo is preferred by the majority of humans.*

MemNet [Khosla et al. 2015] computes a global memorability score for each image. It is trained on the largest annotated image memorability dataset (LaMem) that contains 60,000 images. The network architecture is based on AlexNet. The model is pre-trained on scene and object images and then fine-tuned on LaMem. It is shown that MemNet can reach human consistency in predicting memorability. Intuitively, the preferred image in a series may be the most memorable one, so we also evaluate this model on our dataset. Our results show that MemNet outperforms other previous methods, indicating that memorability has a strong connection with how people choose images in photo series.

Figure 6 shows that both our feature-based methods and end-to-end learning approaches outperform previous methods by both measures. The methods only using ConvNet features perform better than only using handcrafted features, which implies that mid-level features are important in the photo triage problem. Combining both features further improves the performance. While the hybrid method that combines handcrafted features with VGG performs slightly worse than the AlexNet counterpart, the VGG Siamese method achieves the best performance overall – highest likelihood with 73% accuracy.

In Figure 7, we further analyze the performance of the end-to-end model on subsets of data with different levels of human agreement. It shows that as human opinions become more consistent, the performance of our model on both validation and testing sets also increases. We pick five examples of photo pairs, each from one human agreement level, to illustrate some success and failure cases. For Figure 7(b), our predictor does not pick the slightly better one, but human decisions are also widely divided. For pair (d), people prefer the upper one, perhaps due to a small back-facing person in the lower photo; but our method fails to recognize it given the current Siamese network takes low resolution images as input. For pair (f), the upper scene is heavily occluded, but people prefer it because it shows that the game is well-attended. Our model cannot capture this high level, semantic image interpretation. The successful predictions of our model on pairs (c) and (e) show that our model can make correct decisions based on different factors such as composition and color in each specific case.

6 Applications

A direct application of our work is to develop smart album viewers. In addition to providing faster photo triaging, a smarter viewer could automatically identify photo series, and show only the top images for each series instead of presenting all images in it to the user at once. This allows a more efficient interface to manage photo albums and the photo series in them. Beside triage, here we describe a few other photo and video editing applications that can benefit from the proposed dataset and methods.

Automatic Photo Enhancement

We have discovered that in our dataset, there are plenty of image pairs where one photo is strongly preferred over the other for only one major reason, such as composition, color, or exposure. Such pairs are valuable examples to learn not only human preference in each aspect, but also the corresponding transformation for improving human preference that can be applied to a new photo. We present a new example-based photo enhancement framework similar in spirit to Image Analogy [Hertzmann et al. 2001]. We demonstrate two applications of this framework: auto crop and color correction.

We first build a dataset of photo pairs for each application. We collect photo pairs such that for each pair after applying a transformation (color or crop), the less favored photo (B_i for "bad") becomes visually similar to the preferred one (G_i for "good"). Next, given a new photo N, we search through all pairs in the dataset and find the pair i for which B_i is most similar to N (by L_2). Then we can automatically apply a transformation to make an enhanced image E by the analogy $B_i : G_i :: N : E$.

The search for the nearest B_i depends on the operation. For color correction, we concatenate L*a*b* color histogram and content features introduced in Section 4.1. For auto crop, we only consider images B_i that have roughly the same aspect ratio as N, and extract GIST features with 8 blocks (images resized to 256^2) for finding the nearest neighbor. (We also tried content features as we did for color correction, but in our experiments found GIST features perform better for describing the geometry for cropping purposes.)

Next we compute a transformation from the selected photo pair and apply it to N to produce E. For cropping, we (1) use SIFT matching [Lowe 2004] to find a perspective transformation from B_i to G_i; and (2) if this is nearly a similarity transformation, we choose the crop that best approximates it. For color correction, we use the NRDC approach [HaCohen et al. 2011] to find a global color mapping from B_i to G_i. Figure 8 shows two examples of enhancements made by this method, and more examples may be found on our project webpage `phototriage.cs.princeton.edu`.

	Original	Commercial	Ours
Auto Crop	22%	29%	49%
Auto Color	9%	48%	43%

Table 1: *Comparison of our proposed automatic cropping method with AutoCrop feature in Photoshop Elements, and our automatic color enhancement with AutoTone in Lightroom. Humans compared the original and two automatically enhanced photos. Percentages are fraction of photos where a particular variant was most preferred of the three options. Our cropping method performs best. Even though humans prefer the Lightroom enhancement most often, ours is still preferred in many cases.*

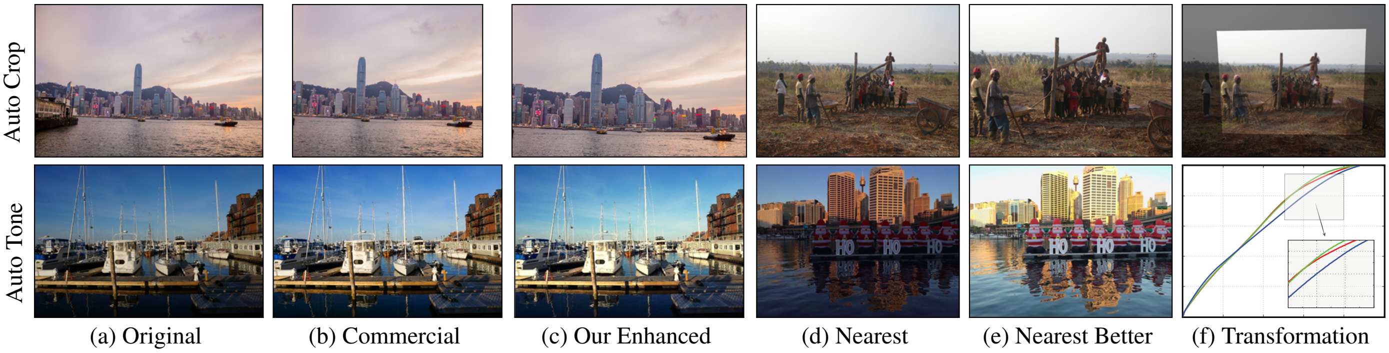


Figure 8: *Automatic enhancement. The original (a) is improved either by a commercial product (b) or our proposed method (c). In the proposed method, a nearest neighbor (d) is found in the dataset where a better alternative (e) from the same series implies a transformation (f) that can be applied to the original (by analogy) to improve it.*

To objectively evaluate the proposed editing framework, we apply it on two new test datasets: one with 468 images for cropping; the other with 556 images for color correction. We conduct additional MTurk user studies to compare our results against the originals, as well as the automatic enhancement results produced by Adobe "Auto Crop" and "Auto Color" features. The results are reported in Table 1, suggesting that our results are in general quite comparable to those produced by existing commercial tools that are dedicated for these tasks. Although the nearest image B_i found in this relatively small dataset often has very different content than the input N, they share similar characteristics in either color or composition. Thus the computed transformation can often produce reasonable results. Nevertheless, we expect the performance of this approach to improve with a larger dataset.

Video Filmstrips

A number of applications rely on a "filmstrip" that summarizes a video clip. These are often shown in video browsing and editing software as well as browsers for video on smart phones. Uniformly sampling frames from the video sometimes chooses good frames and sometimes bad ones, essentially by luck. A line of research seeks to summarize video by various approaches, for example modeling human attention [Ma et al. 2002]. We propose a simple alternative approach based on our learned model of human preference among a series of photos. We extract sequential groups of keyframes from the video, uniformly sampled in time, and choose the "best" from each series using our learned model for preference. This approach is compared with uniform sampling in Figure 9. These examples are about 80 sec, so each frame of the filmstrip represents about 16 secs or 480 raw frames of video (from which our method chose the "best" of around 30 keyframes). In one case the same frame is chosen, but more often than not we find the frames selected from our video are indeed better than the uniformly sampled ones.

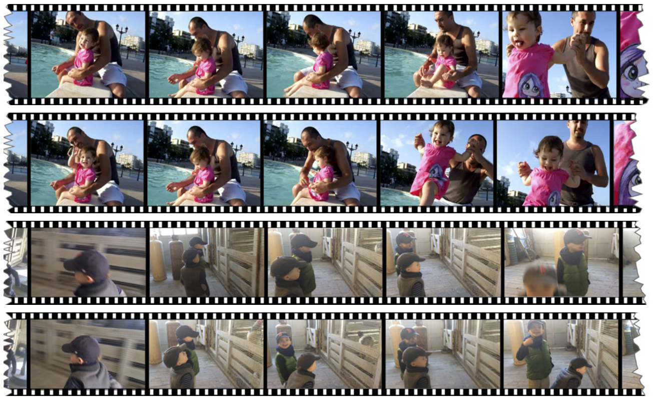


Figure 9: *Video Filmstrips. Uniformly sampling in time (first and third rows) gets a mix of good and bad frames. Our approach (second and fourth row) uses our learner to pick the "best" frame from a region around each uniformly sampled frame. Video frames courtesy of the Flickr users Oleg Sidorenko (upper) and Juhan Sonin (lower).*

7 Discussion and Future work

We study the problem of evaluating relative image quality in a series of photos capturing roughly the same scene, a key step towards automatic, intelligent triage of personal photos. Our contributions include the first large-scale photo series dataset collected from personal photo albums, with quantitative human preference for each image obtained through a massive user study. We propose several learning approaches that outperform previous methods on this task, and further show how the proposed method can be used for other applications such as image cropping and color adjustment.

Although significant progress has been made in this work, it is worth noting that even our best result is still significantly lower than the average human performance. However, we are optimistic that in the foreseeable future, more advanced solutions can be developed to achieve human performance on this task. Here we discuss a few possible directions for future exploration. From the user comments received from the user study, it is clear that people use features at all levels for reasoning about the relative quality when comparing similar images. When low-level and mid-level features are similar, people tend to rely on high level ones such as "interestingness" of the subjects to make a decision. Currently our model only learns low-level and mid-level features, but does not apply semantic level scene understanding, which is an open area for future improvement. Moreover, our current model outputs a binary decision for preference among a pair of photos. It would be useful for many applications to train a regression model that indicates how strongly one image would be preferred over another. However, this problem is more challenging, and remains for future work.

We have also discovered from the user comments that people often look at very different set of features when evaluating photos of different scene categories. For example, in photos that contain people, their appearance strongly dominants the evaluation results. This suggests that a better strategy may be to first classify images into a few distinct categories, then learn a specific model in each category for evaluation. Such categories may be automatically generated by analyzing the user comments, and grouping photos that are associated with similar keywords. Furthermore, our current models only include facial features, but the user comments suggest that other human attributes, such as pose or gaze, are also important features. Including these features may further improve the performance.

Acknowledgements

We thank the anonymous reviewers for helpful suggestions. This work was supported in part by Adobe and the NSF (IIS-1421435). We thank all of the contributors to our photo dataset: Rachel Adler, Doug Ashley, Carolyn Beard, Fei Bo, Guillermo Martinez Cabalga, Ethan Campbell, Anne Caswell-Klein, Adam Cellon, Huiwen Chang, Audrey Chebet, Desai Chen, Stephanie Diu, Rita Fang, Yun Fei, Noah Fishman, Nicolas Freeman, Avigail Gilad, Lisa Gong, Heather Grace, Jesenia Haynes, Sonia Howlett, Joan Hsiao, Emily Hsu, Sohee Hyung, Benjamin Jacobson, Clare Jeong, Jacob Kaplan, Tyler Kaye, Ji-Sung Kim, Tom Kingori, Grace Koh, Jiaming Kong, Taylor Kulp-McDowall, Ashlyn Lackey, Alex Lam, Katherine Lee, Victoria Lin, Kathryn Little, Alex Liu, Cindy Liu, Jennifer Liu, Weber Liu, Annie Lu, Natalie Lu, Angela Mao, Lindsay Martinez, Anna Matlin, Ryan McCaffrey, Mitch Mitchell, John Morone, Shefali Nayak, Anna Pearson, Luke Petruzzi, Selina Pi, Vincent Po, Sahand Keshavarz Rahbar, Andrew Salmons, Leigh Schriever, Paarth Shah, Jessica Shi, Danny Shum, Sarah Spergel, Zachary Stecker, Mark Tengi, Masako Toyoda, Linh Tran, Sarah Tucker, Sadie Van Vranken, Nina Wade, Jue Wang, Kerith Wang, Xue Wang, Ziggie Wang, Samantha Weissman, Jennie Werner, Daniel Wilson, Yihua Xie, Alexander Xu, Dyland Xue, Mandy Yang, Aravind Yeduvaka, Jennifer Yin, Lily Zhang, Maggie Zhang, Mark Zhang, Yinda Zhang, Yuanyuan Zhao, Serena Zheng, Lulu Zhong, Carlos Zhu.

References

BELL, S., AND BALA, K. 2015. Learning visual similarity for product design with convolutional neural networks. *ACM Transactions on Graphics (TOG) 34*, 4, 98.

BHATTACHARYA, S., SUKTHANKAR, R., AND SHAH, M. 2010. A framework for photo-quality assessment and enhancement based on visual aesthetics. In *Proceedings of the international conference on Multimedia*, ACM, 271–280.

BREIMAN, L. 2001. Random forests. *Machine learning 45*, 1, 5–32.

BROMLEY, J., BENTZ, J. W., BOTTOU, L., GUYON, I., LECUN, Y., MOORE, C., SÄCKINGER, E., AND SHAH, R. 1993. Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence 7*, 04, 669–688.

BYCHKOVSKY, V., PARIS, S., CHAN, E., AND DURAND, F. 2011. Learning photographic global tonal adjustment with a database of input/output image pairs. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, IEEE, 97–104.

CAO, X., WEI, Y., WEN, F., AND SUN, J. 2014. Face alignment by explicit shape regression. *International Journal of Computer Vision 107*, 2, 177–190.

CHOPRA, S., HADSELL, R., AND LECUN, Y. 2005. Learning a similarity metric discriminatively, with application to face verification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, IEEE, 539–546.

COOTES, T. F., EDWARDS, G. J., AND TAYLOR, C. J. 1998. Active appearance models. In *Computer Vision?ECCV?98*. Springer, 484–498.

DATTA, R., JOSHI, D., LI, J., AND WANG, J. Z. 2006. Studying aesthetics in photographic images using a computational approach. In *Computer Vision–ECCV 2006*. Springer, 288–301.

DHAR, S., ORDONEZ, V., AND BERG, T. L. 2011. High level describable attributes for predicting aesthetics and interestingness. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, IEEE, 1657–1664.

DRUCKER, S., WONG, C., ROSEWAY, A., GLENNER, S., AND DE MAR, S. 2003. Photo-triage: Rapidly annotating your digital photographs. Tech. rep., Microsoft Research Technical Report, MSR-TR-2003-99.

GIRSHICK, R., DONAHUE, J., DARRELL, T., AND MALIK, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, IEEE, 580–587.

GIRSHICK, R. 2015. Fast r-cnn. *arXiv preprint arXiv:1504.08083*.

GUO, Y., LIU, M., GU, T., AND WANG, W. 2012. Improving photo composition elegantly: Considering image similarity during composition optimization. In *Computer Graphics Forum*, Wiley Online Library, 2193–2202.

HACOHEN, Y., SHECHTMAN, E., GOLDMAN, D. B., AND LISCHINSKI, D. 2011. Non-rigid dense correspondence with applications for image enhancement. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2011) 30*, 4, 70:1–70:9.

HARIHARAN, B., ARBELÁEZ, P., GIRSHICK, R., AND MALIK, J. 2014. Hypercolumns for object segmentation and fine-grained localization. *arXiv preprint arXiv:1411.5752*.

HE, K., ZHANG, X., REN, S., AND SUN, J. 2015. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*.

HERTZMANN, A., JACOBS, C. E., OLIVER, N., CURLESS, B., AND SALESIN, D. H. 2001. Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, ACM, 327–340.

JACOBS, D. E., GOLDMAN, D. B., AND SHECHTMAN, E. 2010. Cosaliency: Where people look when comparing images. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, ACM, 219–228.

JUDD, T., EHINGER, K., DURAND, F., AND TORRALBA, A. 2009. Learning to predict where humans look. In *IEEE International Conference on Computer Vision (ICCV)*, IEEE.

KARAYEV, S., TRENTACOSTE, M., HAN, H., AGARWALA, A., DARRELL, T., HERTZMANN, A., AND WINNEMOELLER, H. 2013. Recognizing image style. *arXiv preprint arXiv:1311.3715*.

KAUFMAN, L., LISCHINSKI, D., AND WERMAN, M. 2012. Content-aware automatic photo enhancement. In *Computer Graphics Forum*, Wiley Online Library, 2528–2540.

KE, Y., TANG, X., AND JING, F. 2006. The design of high-level features for photo quality assessment. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1, IEEE, 419–426.

KHOSLA, A., RAJU, A. S., TORRALBA, A., AND OLIVA, A. 2015. Understanding and predicting image memorability at a large scale. In *International Conference on Computer Vision (ICCV)*.

KITTUR, A., CHI, E. H., AND SUH, B. 2008. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, New York, NY, USA, CHI '08, 453–456.

KRIZHEVSKY, A., SUTSKEVER, I., AND HINTON, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

LIU, L., CHEN, R., WOLF, L., AND COHEN-OR, D. 2010. Optimizing photo composition. *Computer Graphic Forum (Proceedings of Eurographics) 29*, 2, 469–478.

LONG, J., SHELHAMER, E., AND DARRELL, T. 2014. Fully convolutional networks for semantic segmentation. *arXiv preprint arXiv:1411.4038*.

LOWE, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision 60*, 2, 91–110.

LU, X., LIN, Z., JIN, H., YANG, J., AND WANG, J. Z. 2014. Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the ACM International Conference on Multimedia*, ACM, 457–466.

LUO, Y., AND TANG, X. 2008. Photo and video quality evaluation: Focusing on the subject. In *Computer Vision–ECCV 2008*. Springer, 386–399.

LUO, W., WANG, X., AND TANG, X. 2011. Content-based photo quality assessment. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, IEEE, 2206–2213.

MA, Y.-F., LU, L., ZHANG, H.-J., AND LI, M. 2002. A user attention model for video summarization. In *Proceedings of the tenth ACM international conference on Multimedia*, ACM, 533–542.

MARCHESOTTI, L., PERRONNIN, F., LARLUS, D., AND CSURKA, G. 2011. Assessing the aesthetic quality of photographs using generic image descriptors. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, IEEE, 1784–1791.

MEGVII INC., 2013. Face++ research toolkit. www.faceplusplus.com.

MURRAY, N., MARCHESOTTI, L., AND PERRONNIN, F. 2012. Ava: A large-scale database for aesthetic visual analysis. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, IEEE, 2408–2415.

NISHIYAMA, M., OKABE, T., SATO, I., AND SATO, Y. 2011. Aesthetic quality classification of photographs based on color harmony. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, IEEE, 33–40.

OLIVA, A., AND TORRALBA, A. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision 42*, 3, 145–175.

PAIGE, C. C., AND SAUNDERS, M. A. 1982. Lsqr: An algorithm for sparse linear equations and sparse least squares. *ACM Trans. Math. Softw. 8*, 1, 43–71.

PARK, J., LEE, J.-Y., TAI, Y.-W., AND KWEON, I. S. 2012. Modeling photo composition and its application to photo rearrangement. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, IEEE, 2741–2744.

RALPH ALLAN BRADLEY, M. E. T. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika 39*, 3/4, 324–345.

REN, X., AND MALIK, J. 2003. Learning a classification model for segmentation. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, IEEE, 10–17.

REN, S., HE, K., GIRSHICK, R., AND SUN, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 91–99.

REN, S., HE, K., GIRSHICK, R. B., ZHANG, X., AND SUN, J. 2015. Object detection networks on convolutional feature maps. *CoRR abs/1504.06066*.

SIMON, I., SNAVELY, N., AND SEITZ, S. M. 2007. Scene summarization for online image collections. In *ICCV*, IEEE.

SIMONYAN, K., AND ZISSERMAN, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

SINHA, P., MEHROTRA, S., AND JAIN, R. 2011. Summarization of personal photologs using multidimensional content and context. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ACM, 4.

SZEGEDY, C., LIU, W., JIA, Y., SERMANET, P., REED, S., ANGUELOV, D., ERHAN, D., VANHOUCKE, V., AND RABINOVICH, A. 2014. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*.

SZEGEDY, C., VANHOUCKE, V., IOFFE, S., SHLENS, J., AND WOJNA, Z. 2015. Rethinking the inception architecture for computer vision. *arXiv preprint arXiv:1512.00567*.

TANG, H., JOSHI, N., AND KAPOOR, A. 2011. Learning a blind measure of perceptual image quality. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, IEEE, 305–312.

WANG, X.-J., ZHANG, L., AND LIU, C. 2013. Duplicate discovery on 2 billion internet images. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, IEEE, 429–436.

YE, P., KUMAR, J., KANG, L., AND DOERMANN, D. 2012. Unsupervised feature learning framework for no-reference image quality assessment. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, IEEE, 1098–1105.

YU, F., AND KOLTUN, V. 2015. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

YU, F., ZHANG, Y., SONG, S., SEFF, A., AND XIAO, J. 2015. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*.

YUAN, L., AND SUN, J. 2012. Automatic exposure correction of consumer photographs. In *Computer Vision–ECCV 2012*. Springer, 771–785.

ZHANG, L., SONG, M., ZHAO, Q., LIU, X., BU, J., AND CHEN, C. 2013. Probabilistic graphlet transfer for photo cropping. *Image Processing, IEEE Transactions on 22*, 2, 802–815.

ZHOU, E., FAN, H., CAO, Z., JIANG, Y., AND YIN, Q. 2013. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 386–391.

ZHU, J.-Y., AGARWALA, A., EFROS, A. A., SHECHTMAN, E., AND WANG, J. 2014. Mirror mirror: Crowdsourcing better portraits. *ACM Transactions on Graphics (TOG) 33*, 6, 234.