Intelligent Portrait Composition Assistance

— Integrating Deep-learned Models and Photography Idea Retrieval

Farshid Farhat, Mohammad Mahdi Kamani, Sahil Mishra, James Z. Wang
{fuf111,szm5707}@cse.psu.edu,{mqk5591,jwang}@ist.psu.edu

ABSTRACT

Retrieving photography ideas corresponding to a given location facilitates the usage of smart cameras, where there is a high interest among amateurs and enthusiasts to take astonishing photos at anytime and in any location. Existing research captures some aesthetic techniques and retrieves useful feedbacks based on one technique. However, they are restricted to a particular technique and the retrieved results have room to improve as they can be limited to the quality of the query. There is a lack of a holistic framework to capture important aspects of a given scene and give a novice photographer informative feedback to take a better shot in his/her photography adventure. This work proposes an intelligent framework of portrait composition using our deep-learned models and image retrieval methods. A highly-rated web-crawled portrait dataset is exploited for retrieval purposes. Our framework detects and extracts ingredients of a given scene representing as a correlated hierarchical model. It then matches extracted semantics with the dataset of aesthetically composed photos to investigate a ranked list of *photography ideas*, and gradually optimizes the human pose and other artistic aspects of the composed scene supposed to be captured. The conducted user study demonstrates that our approach is more helpful than the other constructed feedback retrieval systems.

KEYWORDS

Photographic Composition; Image Aesthetics; Smart Camera; Portrait Photography; Deep Learning; Image Retrieval.

1 INTRODUCTION

Art still has many ambiguous aspects out of the known sciences, and the beauty of the art is coming from the virgin novelty by artists. It is still daunting for a machine to compose an impressive original song, painting or script. However, high-resolution photography has been made ubiquitous by recent technologies, such as high-quality smart camera phones. Also, the aesthetics of the photography are known as some rules in artistic literature [21, 22, 49] such as balance, geometry, symmetry, the rule of thirds, framing, and etc. Digital photography is of great interest among most people using social networking and photo sharing websites such as Facebook, Google Photos, Twitter, Instagram, etc., but getting a striking photo involves experience and skills, and is often not easy.

MM'17, Mountain View, CA USA

DOI: xx.xxx/xxx_x

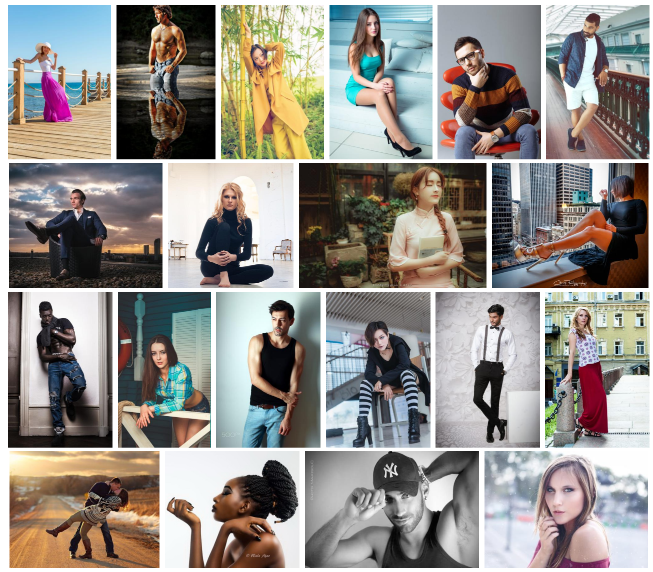

Figure 1: Portrait images given various scenes with several pose ideas for a better composed photo. Images from the 500px website are selected by our framework.

While there are many styles for photography [34, 41, 49] around the world, selecting proper *photography ideas* for a given scene remains a challenging problem, and yet to be fully investigated. The major problem with taking a good portrait photo in a given location is the lack of a local photographer guide conveying us to capture a good portrait pose. In fact, professional photographers usually have expertise and creativity in making good positions intuitively [13, 46, 50]. Through reading books about photography, one can get familiar with some common composition rules such as balancing, framing, the rule of thirds, etc., but it can still be difficult to select and apply techniques for making genuine photos, in a way similar to the gap between reading and writing a novel.

Some basic rules of photography composition inspired from art books [22, 49, 50] have been used by multimedia and vision researchers as aesthetics features for assessment and evaluation of the photos [10, 19, 31, 33, 51]. Other approaches manipulate the taken photo in an online system [3, 4] for auto-composition or re-composition. The techniques include smart cropping [37, 42, 43, 47, 48, 54, 56], warping [6, 30], patch re-arrangement [2, 9, 38], cutting and pasting [4, 56], and seam carving [14, 25], but they can barely help an amateur photographer capture a brilliant photo. More specifically in portrait photography, there are rule-based assessment models [20, 32] using known photography basics to evaluate portraits, and facial assessment models [26–28, 39, 53] exploiting facial features including smile, age, gender, and etc. Perhaps on-site photographic feedback systems [18, 24, 55] can help amateur

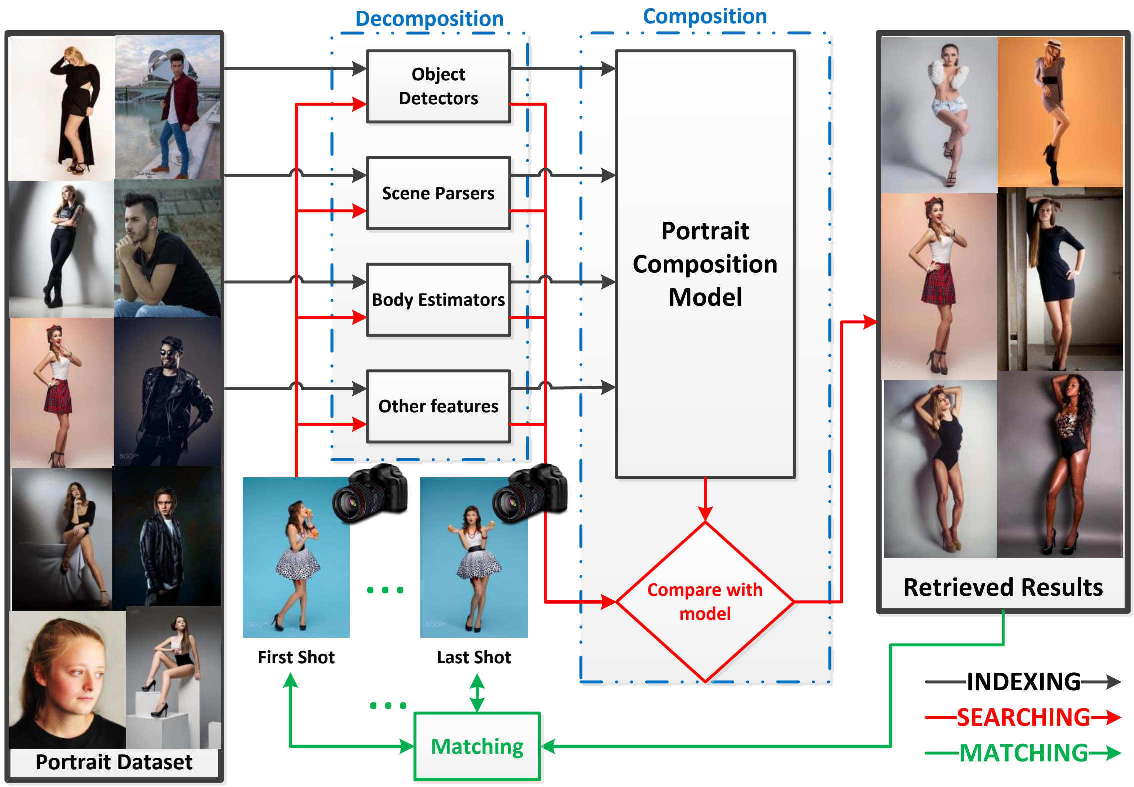

Figure 2: The flowchart of our portrait composition assistance: Black flows show the indexing process, red flows show the searching process, and green flows show the matching process. Decomposition step extracts the image semantics and features, and composition step searches for well-posed images in the dataset based on the semantics and other features.

photographers better by retrieving similar-aesthetic images as a qualitative composition feedback, but their retrieval system is limited to a specific aesthetic feature like perspectiveness [59, 60] or triangle [16].

In this paper, we focus on an assistant framework that helps people make a better pose for their portrait photo with regard to their current location. Given a prior shot from the photographer or the camera viewfinder, our portrait composition assistance outputs some highly-rated prior-composed photos as an assessed feedback. Figure 1 shows some highly-rated portrait images, many taken by professionals, collected from the 500px website and selected by our framework. These 20 portraits are captured in various locations and scenes, and can be divided into categories such as facial, full body, upper body and couple. Each of them has its own photography idea(s) such as a woman with hat (1st image) has made a apropos pose at the heart of the leading lines (fence), or a girl sitting with crossed ankles bended legs (4th image) where this pose creates a nice S-shape. These techniques are believed to make portrait photography more appealing.

Specifically, we address aesthetic retrieval and evaluation of the human poses in portrait photography, and try to improve the quality of the next shot by providing meaningful and constructive feedback to an amateur photographer. Figure 2 shows the flowchart of our approach to assist an amateur photographer in getting a better shot. Based on the first shot as a query, some high-ranked well-posed results are retrieved from our designated dataset using portrait composition model containing the dataset features, and the results are illustrated to the photographer to help compose a better shot, and the last shot is captured when the current pose is matched with one of the results closely. The details of the flowchart has been explained in later sections. The main **contributions** are as follows:

- A holistic framework to intelligently assist amateur photographers to compose a better portrait using our proposed deep-learned models and image retrieval system.
- Various improved deep-learned detectors including object detector, scene parser and pose estimator to extract semantics are integrated.
- Scene construction by composing the semantics and retrieving the desired images from the dataset. We match the scene ingredients with semantic retrievals to optimize the final pose.
- Creating a large dataset containing over 320,000 highly-rated aesthetically composed portraits, with several categories and various scenes.

2 RELATED WORK

General Visual Aesthetics Assessment: While there are many books in art to guide people mastering the challenges of taking professional photographs, the conducted research in technical fields mostly focus on the evaluation and manipulation of the images, after the photo is taken. Basic image aesthetics and composition rules in art [21, 22, 49] as visual semantic features have first been studied computationally by Datta *et al.* [10] and Ke *et al.* [19]. Luo *et al.* [31] and Wong and Low [51] attempt to leverage a saliency map method to focus on the features of the salient part as the more appealing part of the image. Marchesotti *et al.* [33] show that generic image descriptors can be very useful to assess image aesthetics, and build a generic dataset for aesthetics assessment called as Aesthetic Visual Analysis (AVA) [36].

Image Re-Composition: Auto-composition or re-composition systems [3, 4] can passively change the taken photo for a better composition. Cropping techniques [43, 47, 48] separate the region of interest (ROI) by the help of saliency map or eye fixation, basic aesthetic rules [56], or visual aesthetics features in the salient region [37, 42, 54]. As another type of re-composition, Liu *et al.* [30] use warping, *i.e.*, representing the image as a triangular or quad mesh, to map the image into another mesh while keeping the semantics and perspectiveness. Also, R2P [6] detects the foreground part in reference and input image, and tries to re-target the salient part of the image to the best fitted position using a graph-based algorithm. Furthermore, patch re-arrangement techniques patch two ROIs in an image together. Pure patch rearrangement [2, 9, 38] detects the group of pixels on the borders of the patch and matches this group to the other vertical or horizontal group of pixels near the patched area. Cut and paste methods [4, 56] remove the salient part, and re-paint the foreground with respect to salient part and borders, and then paste it to the desired position in the image. Seam carving [14, 25] replaces useless seams.

Portrait Aesthetics Assessment: While there are a lot of works in image aesthetics assessment, a few of them consider portrait photography deeply. Even in this domain, they haven't tried to explore a novel method to solve the problem in photographic portraiture, rather they just combine and use old known features or modified trivial ones to apply in the facial domain. We can categorize them into two main groups: rule-based evaluation models [20, 32] exploit known photography rules to assess portraits, and facial evaluation models [26–28, 39, 53] use visual facial features like smiling, age, gender, etc. Khan and Vogel [20] claim and show a small set of right spatial features can perform better than a large set of aesthetics. Also, their feature importance analysis interestingly shows their spatial features which are not obeying the rule of thirds, mostly affect the system accuracy. Males *et al.* [32] explore headshot aesthetic quality by means of some famous rules, low-level and face-related features. Xue *et al.* [53] study the design inferring portrait aesthetics with more appealing facial features like smiling, orientation, to name but a few. Similarly while exploiting traditional features like Hue, saturation, brightness, contrast, simplicity, sharpness, and the rule of thirds, also their novelty is summed up as extracting saliency map by graph-based visual saliency [15], and calculating standard deviation and main subject coincidence of the saliency map. The other facial evaluation models [26–28] use old known low-level aesthetics features such as colorfulness, sharpness and contrast as well as high-level facial features such as gender, age, and smile. Their idea is based on exploiting these features for all segmented parts of the face including hair, face, eyes, and mouth. Redi *et al.* [39] interestingly show that the beauty of the portrait is related to the amount of art used in it not the subject beauty, age, race, or gender. While using a large dataset from AVA [36], they exploit a high-dimensional feature vector including aesthetics rules, biometrics and demographics features, image quality features, and fuzzy properties. Based on lasso regression output, eyes sharpness and uniqueness have the highest rank to be a good portrait.

Feedback on Photographic System: An aesthetics assessor system may find a metric value to evaluate an input image, but the way it conveys this information to photographer is more crucial, since the photographer probably has no idea about how to improve the aesthetics features of the image. That is why providing meaningful feedback to enhance the future shots and not just aesthetics assessment is our final goal in this work. Giving feedback on a photographic system firstly is mentioned by Joshi *et al.* [18], as they suggest a real-time filter to trace and aesthetically rate the camera shots, and then the photographer retake a better shot. On-site composition and aesthetics feedback system (OSCAR) [24, 55] helps smartphone users improve the quality of their taken photos by retrieving similar-aesthetic images as a qualitative composition feedback. Also it gives color combination feedback for having good colorfulness in the next taken photo, and outputs the overall aesthetics rating of the input photo as well. OSCAR is assumed to fulfill future needs of an amateur photographer, but giving such feedbacks might be unrelated or unrealistic to the user, and also it is restricted to a pretty small database in terms of coverage, diversity and copyright. Xu *et al.* [52] suggest to use three-camera array to enhance the quality of the taken photos by the rule of thirds. In fact, the smartphone interface using the camera array information shows some real-time guideline to the user for taking photo from another position. More recently general aesthetic techniques including perspectiveness [59, 60] and triangle [16] methods are exploited to retrieve proper images as an on-site guidance to amateur photographers, but they are restricted to basic ideas in photography while the pose and scene content are ignored.

3 THE METHOD

In this section, we describe the way that we come up with our proposed framework to assist an amateur photographer intelligently capture beautiful photos from the scenes around. More specifically our proposed approach focuses on scene semantics and aesthetic features in portrait images, while in our opinion our ideas can be extended to genres. The flow of proposed framework (Figure 2) includes indexing, searching, and matching.

3.1 Our Approach

A portrait image not only contains a face but also may contain human body including head, trunk, arms, hands, and feet. Beauty of a portrait depends on the foreground positions of the human parts as well as the constellation of the background objects. The goal of portrait composition assistance is to aid a photographer to capture a better portrait given his or her current photography location. The system input is an amateurishly taken photo by the photographer or an automatically taken photo from camera viewfinder. The system output is a feedback (*e.g.* image, animation, comment, etc.) to guide the photographer to get better shots in next shots. A useful feedback as a side information can be any professional photo taken in a similar scene having a good pose with respect to the location. We name such feedback as a *photography idea* because master photographers usually have their own ideas and each taken photo can be categorized as a new idea.

While most of image aesthetics studies are focusing on image assessment and manipulation of captured photos as mentioned in Section 2, there is a lack of innovative active helping with an amateur photographer to take a better shot. Also, available photographic feedback systems [16, 18, 24, 55, 59, 60] have limitations to filter unrelated photography categories or cover a broad range of photography ideas. For instance, a general retrieval system [18, 24, 55] consists of mixed photography categories including portrait, landscape, closeup, to name but a handful. Hence, this leads to an unrelated output from the input, or a feedback which is limited to a narrow range topic such as perspective photos [59, 60] or photos having triangles [16]. The current available frameworks could not remedy the thirst of the beginners for getting professional-looking snapshots. Also, the more challenging part of the problem is that this treatment is not only a single point but also an ambiguous region because of the subjectivity in art. Expressly, there is no unique solution for an unique input, and based on various unseen tastes and manners of the subject, there may be a range of various related feedbacks.

Our approach is inspired from strategy of professional photography [13, 46, 50] because artists gradually make a subject perfect for the last shot, while they usually have a "to-do" list and a "not-to-do" list in their mind. But the difference is that we do not have access to a studio to compose a new environment for our subject, and some of background objects are static or naturally composed before. For example, when we are in the woods, the trees and sky are invariant with respect to our abilities. However, human bodies, animals, or some objects are posable, dynamic or movable. Furthermore, the number of photography ideas for any given location is not limited to any boundary. Even if we assume that the number of photography ideas in the world are limited, this number would be very high (*e.g.* 10K). To our knowledge, the performance of the deep learning models to classify an idea among high number of correlated ideas degrades substantially. Similarly, there is no accurate food category detector from dish image, because the number of food categories is high (*e.g.* 65K) and the recipe retrieval is done after ingredients detection [8].

Our method includes *decomposition* and then *composition* of semantic objects, and *matching* the proper pose. Like a chess puzzle, we should understand the constellation of the scene, and then move toward the best position. Similarly, we decompose the input shot from the amateur photographer or camera viewfinder into as many as observable objects. Indeed, we extract high level semantics of the scene in the shot, and then realize these semantics as a whole with available photography ideas in the dataset. Up to this step called as *semantic retrieval*, we find the proper photography idea based on the current scene ingredients. In the next step of our methodology known as *matching*, we follow the subject via viewfinder to match his/her pose with the available ideas, and automatically shot the scene similar to "smile shot" mode in smart cameras.

3.2 The Dataset

The most valuable resource of this work is the collected dataset, because it contains a large number of the innovative photography ideas from around the world. To get this done, we tried many photo-sharing websites for photography purposes including Flickr, Photo.net, DPChallenge, Instagram, Google Photos, Pinterest, and Unsplash. However, none of them could properly cover several categories of portrait photography comprising full body, upper body, facial, group, couple or any two, side-view, hand-only, leg-only, to mention but a few. The process of searching, collecting, and updating the dataset is very time-consuming and taxing, hence, automating this process is quite helpful.

Our dataset is constructed by crawling the 500px website which contains photos from millions of creative photographers around the world expanding their social networks of colleagues while exploiting technical and aesthetic skills to make money. To get the JSON file list of the images sorted by rating, we wrote a distributed multi-IP, block-free Python script mostly using keywords including portrait, pose, human, person, woman, man, studio, model, fashion, male, female, and so on. We end up with over 320,000 images dataset where the number of images is still growing.

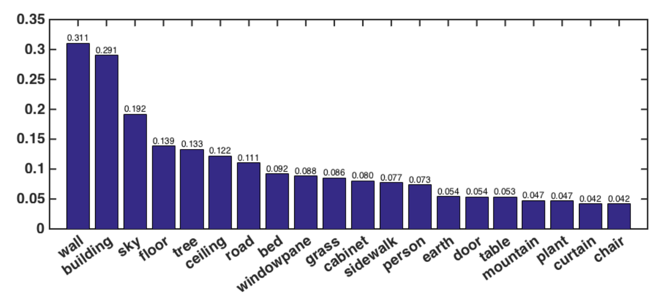

Figure 3: The distribution of the high-repetitive semantics in our dataset.

Finally, we construct a large dataset for photography ideas specially for the above portrait categories (full body, upper body, facial, group, couple or any two, sideview, hand-only, and leg-only) from highly-rated images taken by professional photographers. If we consider the semantics with area greater than the 1% of the image area, we could calculate the probability of the highly-repeated semantics in our dataset (*i.e.* the frequency of the semantic divided by the total number of images). These probabilities are shown in Figure 3, while we have removed "person" (probability=0.9) and "wall" (probability=0.78) from the figure, because they are dominant semantics in most of the images. Definitely having diverse semantics with high frequency in our dataset makes the proposed recommendations with respect to the query shot more helpful. After collecting the dataset and filtering the portraits, we describe the way to retrieve the corresponding results for the query image taken by the camera viewfinder in the following sections.

3.3 Decomposition: Semantics Extraction

Extracting the detected objects in the scene as semantics of the scene is our goal in this section. Then, we construct available scenes in our dataset from the detected semantics and match these semantics with a sub-collection of retrieved photos in our dataset. To achieve this goal, we explain our decomposition strategy which takes the query image from the user and gives a sorted weight list of detected semantics.

While deep learning based models can help computer vision researchers map from nearly unbounded random data domain to a nice classified range, there are still many restrictions to exploit them for applied problems. As mentioned in Section 3, there is *no limit* for artists to create any innovation in art domains such as portrait photography. Hence, it is very difficult if not impossible for available deep learning architectures to learn all of these correlated ideas and do the classification based on the input query with high accuracy. While the number of ideas increases, mean average precision (MAP) falls abruptly with the rate of $O(\frac{1}{n})$. Also manual idea labeling of a huge dataset is costly in terms of time or money.

To tackle the problem of classifying to a large number of ideas, we detect as many objects as possible in the scene instead of photography ideas. In fact, we believe the scene captured in viewfinder consists of various static and dynamic objects. State-of-the-art deep-learned detectors (YOLO [40], PSPNet [57] and RTMPPE [5]) are customized for our purpose. YOLO [40] neural network trained on MSCOCO dataset [29] partitions the query photo into several bounding boxes predicting their probabilities. Pyramid scene parsing network (PSPNet) [57] as the winner of scene parsing challenge on ADE20K dataset [58] uses global context information through a pyramid pooling module. PSPNet predicts the scene objects in pixel-level. Real-time multi-person 2D pose estimation (RTMPPE) predicts vector fields to represent the associative locations of the anatomical parts via two sequential prediction process exposing the part confidence maps and the vector fields on MSCOCO [29] and MPII [1] datasets. To improve the accuracy, we have re-trained YOLO, PSPNet, and RTMPPE models on extended MSCOCO and ADE20K datasets by adding some of failed cases from our 500px dataset as an augmented training dataset. The illustration of some sample results are in Figure 4, where YOLO object names are shown in a red rectangle with a probability, RTMPPE pose is shown as a colorful connection of skeleton joints, and PSPNet scenes are colorized pixel-wisely based on the pixel codename.

We unify the outputs of the detectors in terms of pixel-level tensors, *i.e.*, our modified YOLO outputs MSCOCO object IDs among 80 categories (from 1 to 80) and their scores as the minus logarithm of their NOT probability ($-\log(1-p)$) for each pixel of the image is representing as a tensor. Also our version of PSPNet outputs ADE20K object IDs among 150 categories (from 1 to 150) and the score for each pixel of the image is represented as a tensor. Similarly, our version of RTMPPE gives 18 anatomical part IDs with their scores as a tensor. So, for any image ($I_{m\times n}$) we have:

$$T^{I,od}_{m\times n\times 2} = \left[t^{I,od}_{i,j,k}\right]\ ,\ t^{I,od}_{i,j,1} = C^{I,id}_{i,j}\ ,\ t^{I,od}_{i,j,2} = -\log_2(1-p^{I,od}_{i,j})\ ,$$
$$T^{I,sp}_{m\times n\times 2} = \left[t^{I,sp}_{i,j,k}\right]\ ,\ t^{I,sp}_{i,j,1} = A^{I,id}_{i,j}\ ,\ t^{I,sp}_{i,j,2} = -\log_2(1-p^{I,sp}_{i,j})\ ,$$
$$T^{I,pe}_{m\times n\times 2} = \left[t^{I,pe}_{i,j,k}\right]\ ,\ t^{I,pe}_{i,j,1} = J^{I,id}_{i,j}\ ,\ t^{I,pe}_{i,j,2} = -\log_2(1-p^{I,pe}_{i,j})\ ,$$

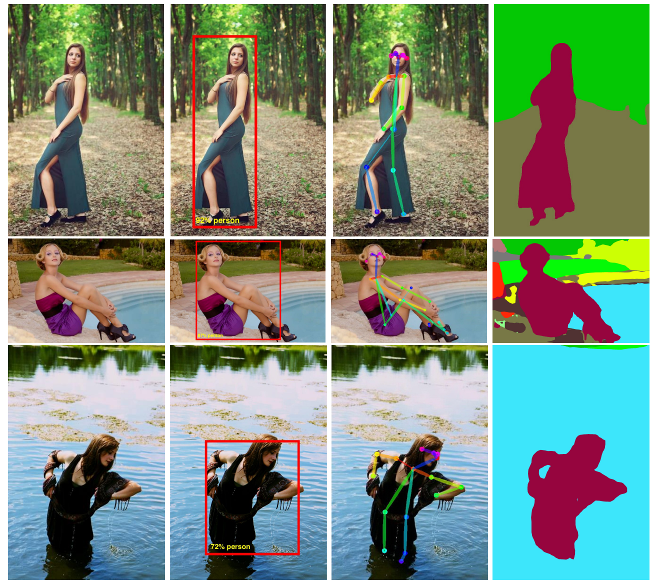

Figure 4: Sample results of our dataset where respectively from left to right are: original thumbnail, YOLO, RTMPPE, and PSPNet illustrations.

where I is an input image, m is the number of rows, n is the number of columns in the image, $T^{I,od}$ is corresponding tensor of object detector (*e.g.* YOLO), $C^{I,id}_{i,j} \in \{1..80\}$ is MSCOCO ID of the pixel at (i,j), $p^{I,od}_{i,j}$ is the MSCOCO ID probability of the pixel at (i,j); $T^{I,sp}$ is tensor of scene parser (*e.g.* PSPNet), $A^{I,id}_{i,j} \in \{1..150\}$ is ADE20K ID of the pixel at (i,j), $p^{I,sp}_{i,j}$ is the ADE20K ID probability of the pixel at (i,j); $T^{I,pe}$ is tensor of pose estimator (*e.g.* RTMPPE), $J^{I,id}_{i,j} \in \{1..18\}$ is the joint ID of the pixel at (i,j), and $p^{I,pe}_{i,j}$ is the joint ID probability of the pixel at (i,j).

To auto-tag or auto-label the image, we integrate these unified results in terms of the objects, their coordinates, and their scores (or probabilities). The number of the detectable objects is 210 objects by merging MSCOCO (80 categories) and ADE20K (150 categories) objects and deduplicating 20 objects. Also we have 18 joints from RTMPPE including nose, neck, right shoulder, right elbow, right wrist, left shoulder, left elbow, left wrist, right hip, right knee, right ankle, left hip, left knee, left ankle, left eye, right eye, left ear, and right ear. YOLO detection for full-body small persons in the image is poor, but it can detect big limbs of the body as a person well. RTMPPE detection for occluded bodies is poor but the detection for full-body persons is acceptable. Also PSPNet detection for objects, not persons, is relatively good compared to others.

First, we detect any person in the image. Our detector's integration scheme has LOW and HIGH thresholds for each detector. These thresholds are trained by a random set of highly-rated ground-truth portraits. If the average score of the pixel with person/limb ID in the image is higher than its HIGH threshold, there is a person in the image, otherwise if the average score of the pixels with person/limb ID in the image is lower than the corresponding LOW threshold, the image will be ignored from indexing or searching process, and we wait for another image for indexing or another shot for searching.

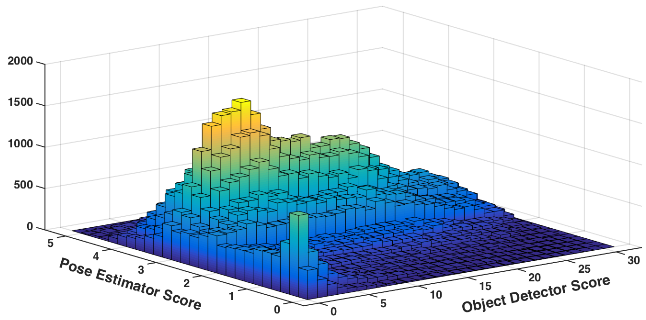

Figure 5: The 3D histogram of the portrait images binned by the object detector and pose estimator scores.

We called this detector's integration scheme as *Hysteresis* detection. Actually it is assured that the confidence ratio of the person in the image with his/her limbs is in a good condition using Hysteresis detection. Using this filtering on our dataset, about 90% (280K+) of the images are passed. The 3D histogram of our portrait dataset in Figure 5 illustrates the frequency of the images smart-binned by the normalized object detector and pose estimator scores. In fact, it shows the effectiveness of integrating the detectors to capture the usefulness of the dataset images more precisely, because we are unifying the detection results for more broad range of ideas not intersecting them to have more confident narrow range of ideas.

Second, we estimate the types of the portrait in the image respectively as two (couple or any two persons), group (more than two persons), full-body, upper-body, facial, sideview, faceless, headless, hand-only, and leg-only. The number of persons is estimated by the max number of person IDs higher than their corresponding HIGH thresholds via YOLO and RTMPPE models. Otherwise, if the image has a nose, two eyes, a hand and a leg OR a nose, an eye, two hands and two legs, it will be categorized as full body. Such combinations are learned after trying some random images as ground truth, because RTMPPE model is not perfect and also in some cases, the limbs are occluded by the others. After examining full-body, if the image has a nose and two eyes and one hand will be divided as upper-body. After category indexing of our dataset, the distribution of the portrait categories with respect to the number of corresponding images in each category by total number of images from previous step is shown in Figure 6. Consequently, the number of images for some categories like full-body, upper-body, facial, group, two, and sideview are adequate.

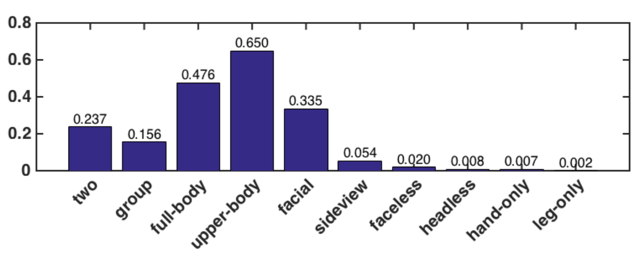

Figure 6: The distribution of the portrait categories with respect to the number of corresponding images.

Third, we go for the other semantics. Some of them are coming from YOLO model (80 categories) and the others are coming from PSPNet (150 categories). At most there are 210 semantics (including person). To rank the order of the semantics, we exploit the max score ID multiply by the saliency map (S) features [17] with our centric distance (D) feature to get our weighted saliency map (W).

$$W(i,j) = \max\left(T^{I,od}_{*,*,2}, T^{I,sp}_{*,*,2}\right) \times S(i,j) \times D(i,j)\,, \quad (1)$$

$$D(i,j) = 1/K \times e^{-\|[i,j]-c\|_k}\,, \quad (2)$$

$$c = \frac{\sum_{i,j} S(i,j).[i,j]}{\sum_{i,j} S(i,j)}\,, \quad (3)$$

where $W(i,j)$ is our weighted saliency ID map, *max* operation is on the 2nd matrix (score matrix) of the tensors, $S(i,j)$ is the saliency map in [17], and $D(i,j)$ is our centric distance feature, K is a tunable constant, c is the center of mass coordinate, and $\|.\|_k$ is the k-th norm operator where $k=1$ in our experiments. Based on various sorted semantics called as portrait scenes, we have indexed our portrait-categorized dataset from the previous step, and Figure 7 depicts the number of portrait scenes for some of highly-frequent portrait categories.

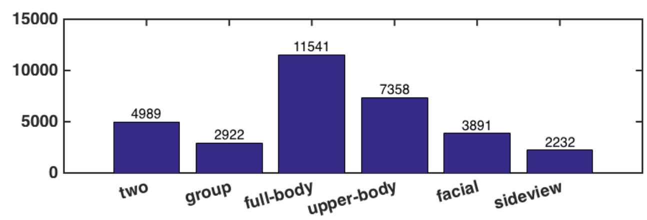

Figure 7: The frequency of the portrait scenes with respect to the highly-requested portrait categories.

Our weighted saliency map can make the detected objects in order, as we can sort the summation of the scores from the semantics and sort them based on their accumulated weights. The output of this step is an ordered list of detected semantics in the query image. In the next step, we will find the closest image to this auto-tagged ordered object list of the query image.

3.4 Composition: Scene Construction

The goal of composition step is to build up a new scene consisting of the desired objects in it. The input of this stage is the ordered weight list of the semantics as well as the image query. The output of this stage will be a bunch of well-posed images corresponding to the query image. As we focus on portrait images, we desire the targeted image contains a well-posed portrait with similar semantics. That the person is interacting with the objects around is important, because the proposed pose by the system is dependent on them.

As we have collected a dataset containing pretty well-posed portrait, we should dig into the dataset and look for an image with similar object constellation, and the existence of this professional dataset makes us pretty sure that the retrieved photos contain good aesthetic photography ideas. Our image retrieval system is not supposed to find images with similar colors, patterns, or poses but it tries to find images with better poses with similar semantics. So

Algorithm 1 Semantic Retrieval

procedure $SemanticRetrieval(Q \in ViewFinder)$
  $T^{Q,od}_{m \times n \times 2} \leftarrow Object\_Detect(Q)$
  $T^{Q,sp}_{m \times n \times 2} \leftarrow Scene\_Parse(Q)$
  $T^{Q,pe}_{m \times n \times 2} \leftarrow Pose\_Estimate(Q)$
  $T^{Q,pe}_{m \times n \times 2} \leftarrow Common\_Features\_Extract(Q)$
  $c^Q \leftarrow \frac{\sum_{i,j} \|[i,j]-[0,0]\|_k \times S(i,j)}{\sum_{i,j} \|[i,j]-[0,0]\|_k}$
  $D^Q(i,j) \leftarrow 1/K \times e^{-\|[i,j]-c^Q\|_k}$
  $W^Q(i,j) \leftarrow \max\left(T^{I,od}_{*,*,2}, T^{I,sp}_{*,*,2}\right) \times S^Q(i,j) \times D^Q(i,j)$
  $V^{Q,pref} \leftarrow M_{PCM}.W^Q$
  $Retrieved\_Indexes \leftarrow Index\_Sort(V^{Q,pref})$
  $Show\_Top4(Retrieved\_Indexes)$
end procedure

the location of the movable objects doesn't matter, but the detected objects are important.

To look for a semantically composed version regarding to the query, we exploit the ordered weight list of the detected objects in Eq. 1 as well as the other common feature vectors [44] of the query image, since we do not trust that the query image taken by an amateur photographer to be well-posed enough to be the query base for our image retrieval system. In fact, we just want to understand the location around the subject, and then based on the scene ingredients, a well-posed image taken by a professional will be proposed to the photographer.

Based on the ordered weight list of the detected objects in the image, we can do the same operations on all of our images in the dataset, and then retrieve the highest ranked candidates as the results. The operations are the same as mentioned in decomposition section including features from YOLO, PSPNet, and RTMPPE to detect the possible persons, their joints, and other objects with their scores. Also we should compute our weighted saliency map to rank the detected objects. The ordered weight list of the semantics with other common features of our portrait dataset, known as *portrait composition model (PCM)* in Figure 1, is represented as a huge "number of images by number of features" matrix (M_{PCM}), and similarly the ordered weight list and other common features of the image query is represented as a long feature vector (W^I). The distance of the W^I from each row of M_{PCM} is defined as inner vector product. Consequently, we have:

$$V^{pref} = M_{PCM}.W^I \ , \tag{4}$$

where V^{pref} is our preference vector, and if we sort it based on the vector values, the index of the rows represent the highest ranked candidates as our feedbacks to the image query (I). The whole process of semantic retrieval for an input image query (Q) has been shown in Algorithm 1. Our experimental results in later section shows the quality of the results.

Notes on Indexing and Searching: Our semantic retrieval system is equivalent to the flows of indexing and searching in Figure 2. Practically, there are many challenges in image retrieval systems [11, 23, 45] as well as in our case. To improve the speed of our image retrieval system, we compute the decomposition step for all images in our dataset. Indexing procedure is very lengthy for the first time, but at the time of update is fast enough because the semantic and feature extraction for an image is real-time using GPU. Furthermore, indexing procedure for our retrieval system organizes the dataset images into categorized folders labeled by the sorted semantic list. Consequently, the composition step is also fast, as it just extracts the query image ordered semantic list, and then will find the target folder containing similar images, and finally will retrieve first five best results from the folder with respect to the top-4 indexes in sorted preference vector (Eq. 4) explained in Algorithm 1. As we just include semantics with normalized score higher than 10 percentages of the total semantics score, the number of ordered semantics are limited (typically less than 5 names). Also naturally all semantic combinations are not possible, so basically the number of scene semantics are limited.

3.5 Matching

Professional photographer starts to pose the subject from head to toe step by step, while there are many to-do list and not-to-do list for portrait photography in his/her mind. We want to create the same environment in a smart camera to accompany an amateur photographer gradually to his/her perfect shot. From semantic retrieval section, we retrieve the proper photography ideas given the query image of the camera viewfinder, and we assume that the photographer has chosen some of the retrieved images as a desired set, and forgets the others as an ignored set. Now in this section, we explain how to capture the proper pose of the subject in the scene, and trigger the "pose shot" for the camera.

The variant component in our framework is the human pose. The relative positions of the human body parts (including nose, eyes, ears, neck, shoulders, elbows, wrists, hips, knees, and ankles) with respect to the nose position as portrait origin are consisting our pose model. Preferably, we would like to start from the position of the nose ($J_0 = (0,0)$) that is connected to neck (J_1), right eye (J_2), and left eye (J_3) are connected to right ear (J_4), left ear (J_5) as they are on a plane of the head. Also, shoulders (J_6 and J_7) can be recognized by a length and an angle from neck, and similarly elbows (J_8 and J_9) from shoulders, wrists (J_{10} and J_{11}) from elbows, hips (J_{12} and J_{13}) from neck, knees (J_{14} and J_{15}) from hips, and ankles (J_{16} and J_{17}) from knees, *i.e.* they are connected as follows:

$$Pre\,(J_i) = J_0 \ , \ i \in \{0,1,2,3\} \ , \tag{5}$$

$$Pre\,(J_i) = J_1 \ , \ i \in \{6,7,12,13\} \ , \tag{6}$$

$$Pre\,(J_i) = J_{i-2} \ , \ i \in \{4,5,8,9,10,11,14,15,16,17\} \ . \tag{7}$$

So, we can always calculate the absolute position using 2D polar coordinates as follows:

$$J_i = J_j + r_{i,j}.e^{\mathrm{i}\theta_{i,j}} \ , \ i \in \{0..17\} \ , \tag{8}$$

where $j = Pre(i)$ *i.e.* part j is the previous part connected to part i, $r_{i,j}$ is the length from joint J_i to joint J_j, $\theta_{i,j}$ is the angle between the line from joint J_i to joint J_j and the line from joint J_j to joint $Pre(J_j)$, and the line crossing J_0 is the image horizon. i is the unit imaginary number. Note that for a 2D human body $r_{i,j}; \forall i,j$ are fixed, but $\theta_{i,j}; \forall i,j$ can be changed to some fixed not arbitrary extents. Also having 3D pose-annotated/estimated single depth images, similarly we can calculate the relative 3D position of the joints using spherical coordinates. So, we have such action boundaries for joints as follows:

$$\theta_{i,j}^{min} \leq \theta_{i,j} \leq \theta_{i,j}^{max} \ , \ j = Pre(i) \ , \tag{9}$$

$$\phi_{i,j}^{min} \leq \phi_{i,j} \leq \phi_{i,j}^{max} \ , \ j = Pre(i) \ . \tag{10}$$

As a result, a human body pose (J) is represented by:

$$\mathrm{J}^{\mathrm{k}} = \left(J_1^k, J_2^k, ..., J_{17}^k\right) \ , \tag{11}$$

where J^{k} is the pose for k-th person (or k-th image with one person), and $\forall i \in \{1..17\} : J_i^k$ is the i-th coordinate of the k-th person. Also we need a distance metric to calculate the difference between two pose features. So we define the distance metric as follows:

$$D\left(\mathrm{J}^{\mathrm{k}}, \mathrm{J}^{\mathrm{l}}\right) \doteq \sum_{i=1}^{17} \|J_i^k - J_i^l\|_q \ , \tag{12}$$

where $D\left(.\right)$ is the distance operator, where J^{k} is the pose feature for k-th person (or k-th image with one person), $\forall i \in \{1..17\} : J_i^k$ is the i-th coordinate of the k-th person, and $\|.\|_q$ (usually L1-norm or L2-norm) is the $L_q - norm$ function of two equal-length tuples.

Now, the camera viewfinder may take and hold several photos gradually from the scene, and finally choose the best among them to save on the camera disk. Actually our matching algorithm searches among the taken photos to get the nearest pose to one of the collected ideas. It is an integer programming problem to find the best seed among all of photography ideas. Given the distance operator of two pose features explored in 12, we can construct our optimization problem as the maximum over all taken photos of the difference of the minimum distance of the ignored set and minimum distance of the desired set. Mathematically, we compute the following optimization problem subject to 9 and 10:

$$I_w = \arg \max_{\forall I_i \in I^t} \left(\min_{\forall Q_j^g \in Q^g} D\left(\mathrm{J}^{\mathrm{Q_j^g}}, \mathrm{J}^{\mathrm{I_i}}\right) - \min_{\forall Q_k^d \in Q^d} D\left(\mathrm{J}^{\mathrm{Q_k^d}}, \mathrm{J}^{\mathrm{I_i}}\right) \right) \ ,$$

where I_w is the wish image, I^t is the set of taken photos, Q^g is the set of ignored retrieved ideas, Q^d is the set of desired retrieved ideas, $D\left(.\right)$ is the distance operator in 12, J^{x} is the pose for x-th image with one person in 11. The optimization problem in continuous mode (not over taken images set) may have (a) solution(s) in feasible region, and in L1-norm case, it is equivalent to multiple linear programming problems but the complexity of the problem will be exponential, and also the solution is not always a desired pose.

3.6 User Study

Currently, there is no other similar or comparable system in the literature to compare with our proposed framework. To evaluate the functionality and the performance of our method, and measure how much the recommended photos make sense and is helpful to the photographer, we conduct a quantitative user study based on the human subject results to compare our method with state-of-the-art semantic and scene type retrievals based on CNN [44] and KNN-SVM [35]. We select a variety of image queries based on many types of portrait categories such as background scene and semantics, single versus group, full-body, upper-body, facial, standing versus sitting, male versus female. All 4096 generic descriptors via public CNN model [7] trained on ImageNet [12] are extracted for our huge dataset images as well as the features of KNN-SVM-based method [35]. Using a PHP-based website with a usage guidance, the hypothesis tests are asked, and the outputs of the methods are randomly shown in each row to fairly be chosen by more than thirty participants among graduate students. Finally, our framework received 65.3% of the 1st ranks among the tests compared to 27.1% CNN as 2nd rank and 7.6% KNN-SVM as 3rd rank. Figure 8 illustrates the results of all methods (CNN: 1st, KNN-SVM: 2nd, ours: 3rd row) with respect to a similar shot at the left side. As it is realized from Figure 8 and our study, the other methods cannot capture portrait categories, scene structure, and corresponding poses of the query shot very well, because the badly-posed full-body query shot is suggested as upper-body, facial, and back poses by other category-agnostic methods. As we hierarchically index our dataset by portrait images, portrait categories, and then semantic categories, our semantic-aware framework accessing to our indexed dataset can retrieve related photography ideas.

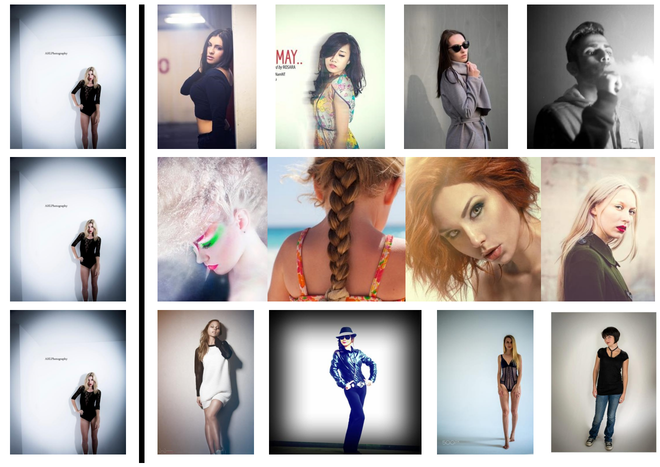

Figure 8: The results of CNN (1st row), KNN-SVM (2nd row), and our method (3rd row) for a sample shot at the left side.

4 CONCLUSIONS

Not only do we collect a huge dataset for portrait photography we also introduce a new framework for portrait composition assistance which aids amateur photographer to capture a better shot from his/her human subject. As the number of photography ideas are increasingly high, directly retrieving and matching the viewfinder photo with an image in our dataset is complicated. Furthermore, the retrieving system not only finds similar images but also searches for images with similar semantics through decomposition and composition stages. After turning in the feedbacks for the photographer, the camera tries to match the final pose with one of the retrieved feedbacks, and make a pose-shot. The performance of our framework has been evaluated by a user study. Another merit of this work is the integration of the deep-based detectors which can make the whole process automatic not manual. Furthermore, this work can be extended to other photography domains using other appropriate detectors.

REFERENCES

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2014. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*.
[2] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan Goldman. 2009. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics* 28, 3 (2009), 1–24.
[3] Subhabrata Bhattacharya, Rahul Sukthankar, and Mubarak Shah. 2010. A framework for photo-quality assessment and enhancement based on visual aesthetics. In *ACM Conference on Multimedia*. 271–280.
[4] Subhabrata Bhattacharya, Rahul Sukthankar, and Mubarak Shah. 2011. A holistic approach to aesthetic enhancement of photographs. *ACM Transactions on Multimedia Computing, Communications, and Applications* 7, 1 (2011), 21.
[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2016. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *arXiv preprint arXiv:1611.08050* (2016).
[6] Hui-Tang Chang, Po-Cheng Pan, Yu-Chiang Frank Wang, and Ming-Syan Chen. 2015. R2P: Recomposition and Retargeting of Photographic Images. In *ACM Conference on Multimedia Conference*. 927–930.
[7] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531* (2014).
[8] Jingjing Chen and Chong-Wah Ngo. 2016. Deep-based ingredient recognition for cooking recipe retrieval. In *ACM on Multimedia Conference*. 32–41.
[9] Taeg Sang Cho, Moshe Butman, Shai Avidan, and William T Freeman. 2008. The patch transform and its applications to image editing. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1–8.
[10] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Wang. 2006. Studying aesthetics in photographic images using a computational approach. *European Conference on Computer Vision* (2006), 288–301.
[11] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. 2008. Image retrieval: Ideas, influences, and trends of the new age. *Comput. Surveys* 40, 2 (2008), 5.
[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*. 248–255.
[13] Christopher Grey. 2014. *Master lighting guide for portrait photographers*. Amherst Media.
[14] YW Guo, M Liu, TT Gu, and WP Wang. 2012. Improving photo composition elegantly: Considering image similarity during composition optimization. In *Computer Graphics Forum*, Vol. 31. Wiley Online Library, 2193–2202.
[15] Jonathan Harel, Christof Koch, and Pietro Perona. 2006. Graph-based visual saliency. In *Advances in Neural Information Processing Systems*. 545–552.
[16] Siqiong He, Zihan Zhou, Farshid Farhat, and James Wang. 2017. Discovering Triangles in Portraits for Supporting Photographic Creation. *IEEE Transactions on Multimedia* (2017).
[17] Laurent Itti, Christof Koch, and Ernst Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 11 (1998), 1254–1259.
[18] Dhiraj Joshi, Ritendra Datta, Elena Fedorovskaya, Quang-Tuan Luong, James Z Wang, Jia Li, and Jiebo Luo. 2011. Aesthetics and emotions in images. *IEEE Signal Processing Magazine* 28, 5 (2011), 94–115.
[19] Yan Ke, Xiaoou Tang, and Feng Jing. 2006. The design of high-level features for photo quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 1. 419–426.
[20] Shehroz S Khan and Daniel Vogel. 2012. Evaluating visual aesthetics in photographic portraiture. In *Symposium on Computational Aesthetics in Graphics, Visualization, and Imaging*. Eurographics Association, 55–62.
[21] Bert Krages. 2012. *Photography: the art of composition*. Skyhorse Publishing, Inc.
[22] David A Lauer and Stephen Pentak. 2011. *Design basics*. Wadsworth Publishing.
[23] Michael S Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain. 2006. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications* 2, 1 (2006), 1–19.
[24] Jia Li, Lei Yao, and James Z Wang. 2015. Photo Composition Feedback and Enhancement. In *Mobile Cloud Visual Media Computing*. Springer, 113–144.
[25] Ke Li, Bo Yan, Jun Li, and Aditi Majumder. 2015. Seam carving based aesthetics enhancement for photos. *Signal Processing: Image Communication* 39 (2015), 509–516.
[26] Arnaud Lienhard, Patricia Ladret, and Alice Caplier. 2015. How to predict the global instantaneous feeling induced by a facial picture? *Signal Processing: Image Communication* 39 (2015), 473–486.
[27] Arnaud Lienhard, Patricia Ladret, and Alice Caplier. 2015. Low level features for quality assessment of facial images. In *Conference on Computer Vision Theory and Applications*. 545–552.
[28] Arnaud Lienhard, Marion Reinhard, Alice Caplier, and Patricia Ladret. 2014. Photo rating of facial pictures based on image segmentation. In *IEEE Conference on Computer Vision Theory and Applications*, Vol. 2. 329–336.
[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*. Springer, 740–755.
[30] Ligang Liu, Yong Jin, and Qingbiao Wu. 2010. Realtime Aesthetic Image Retargeting.. In *Computational Aesthetics*. 1–8.
[31] Yiwen Luo and Xiaoou Tang. 2008. Photo and video quality evaluation: Focusing on the subject. In *European Conference on Computer Vision*. Springer, 386–399.
[32] Matija Males, Adam Hedi, and Mislav Grgic. 2013. Aesthetic quality assessment of headshots. In *International Symposium ELMAR*. 89–92.
[33] Luca Marchesotti, Florent Perronnin, Diane Larlus, and Gabriela Csurka. 2011. Assessing the aesthetic quality of photographs using generic image descriptors. In *IEEE Conference on Computer Vision*. 1784–1791.
[34] Bamberg Matthew. 2010. *101 Quick and Easy Ideas Taken form the Master Photographers of the Twentieth Century*. Course Tecnology CENGAGE Learning.
[35] Joani Mitro. 2016. Content-based image retrieval tutorial. *arXiv preprint arXiv:1608.03811* (2016).
[36] Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *Conference on Computer Vision and Pattern Recognition*. 2408–2415.
[37] Jaesik Park, Joon-Young Lee, Yu-Wing Tai, and In So Kweon. 2012. Modeling photo composition and its application to photo re-arrangement. In *IEEE Conference on Image Processing*. 2741–2744.
[38] Yael Pritch, Eitam Kav-Venaki, and Shmuel Peleg. 2009. Shift-map image editing.. In *International Conference on Computer Vision*, Vol. 9. 151–158.
[39] Miriam Redi, Nikhil Rasiwasia, Gaurav Aggarwal, and Alejandro Jaimes. 2015. The beauty of capturing faces: Rating the quality of digital portraits. In *IEEE Conference on Automatic Face and Gesture Recognition*, Vol. 1. 1–8.
[40] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. 779–788.
[41] Patrick Rice. 2006. *Master Guide for Professional Photographers*. Amherst Media.
[42] A Samii, R Měch, and Z Lin. 2015. Data-Driven Automatic Cropping Using Semantic Composition Search. In *Computer Graphics Forum*, Vol. 34. Wiley Online Library, 141–151.
[43] Anthony Santella, Maneesh Agrawala, Doug DeCarlo, David Salesin, and Michael Cohen. 2006. Gaze-based interaction for semi-automatic photo cropping. In *ACM SIGCHI Conference on Human Factors in Computing Systems*. 771–780.
[44] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN features off-the-shelf: an astounding baseline for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 806–813.
[45] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. 2000. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 12 (2000), 1349–1380.
[46] Jeff Smith. 2012. *Posing for Portrait Photography: A Head-To-Toe Guide for Digital Photographers*. Amherst Media.
[47] Fred Stentiford. 2007. Attention based auto image cropping. In *ICVS Workshop on Computational Attention & Applications*, Vol. 1. Citeseer.
[48] Bongwon Suh, Haibin Ling, Benjamin B Bederson, and David W Jacobs. 2003. Automatic thumbnail cropping and its effectiveness. In *ACM symposium on User Interface Software and Technology*. 95–104.
[49] Roberto Valenzuela. 2012. *Picture Perfect Practice: A Self-Training Guide to Mastering the Challenges of Taking World-Class Photographs*. New Riders.
[50] Roberto Valenzuela. 2014. *Picture Perfect Posing: Practicing the Art of Posing for Photographers and Models*. New Riders.
[51] Lai-Kuan Wong and Kok-Lim Low. 2009. Saliency-enhanced image aesthetics class prediction. In *IEEE Conference on Image Processing*. 997–1000.
[52] Yan Xu, Joshua Ratcliff, James Scovell, Gheric Speiginer, and Ronald Azuma. 2015. Real-time Guidance Camera Interface to Enhance Photo Aesthetic Quality. In *ACM Conference on Human Factors in Computing Systems*. 1183–1186.
[53] Shao-Fu Xue, Hongying Tang, Dan Tretter, Qian Lin, and Jan Allebach. 2013. Feature design for aesthetic inference on photos with faces. In *IEEE Conference on Image Processing*. 2689–2693.
[54] Jianzhou Yan, Stephen Lin, Sing Kang, and Xiaoou Tang. 2013. Learning the change for automatic image cropping. In *IEEE Conference on Computer Vision and Pattern Recognition*. 971–978.
[55] Lei Yao, Poonam Suryanarayan, Mu Qiao, James Z Wang, and Jia Li. 2012. Oscar: On-site composition and aesthetics feedback through exemplars for photographers. *International Journal of Computer Vision* 96, 3 (2012), 353–383.
[56] Mingju Zhang, Lei Zhang, Yanfeng Sun, Lin Feng, and Weiying Ma. 2005. Auto cropping for digital photographs. In *IEEE Conference on Multimedia and Expo*.
[57] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2016. Pyramid Scene Parsing Network. *arXiv preprint arXiv:1612.01105* (2016).
[58] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene Parsing through ADE20K Dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*.
[59] Zihan Zhou, Farshid Farhat, and James Wang. 2017. Detecting Dominant Vanishing Points in Natural Scenes with Application to Composition-Sensitive Image

Retrieval. *IEEE Transactions on Multimedia* (2017).

[60] Zihan Zhou, Farshid Farhat, and James Z Wang. 2016. Detecting Vanishing Points in Natural Scenes with Application in Photo Composition Analysis. *arXiv preprint arXiv:1608.04267* (2016).