ASSESSING PHOTO QUALITY WITH GEO-CONTEXT AND CROWDSOURCED PHOTOS

Wenyuan Yin †, *Tao Mei* ‡, *Chang Wen Chen* †

† State University of New York at Buffalo, NY, USA
‡ Microsoft Research Asia, Beijing, P. R. China

ABSTRACT

Automatic photo quality assessment emerged as a hot topic in recent years for its potential in numerous applications. Most existing approaches to photo quality assessment have predominantly focused on image content itself, while ignoring various contexts such as the associated geo-location and timestamp. However, such a universal aesthetic assessment model may not work well with significantly different contexts, since the photography rules are always scene and context dependent. In real cases, professional photographers use different photography knowledge when shooting various scenes in different places. Motivated by this observation, we leverage the geo-context information associated with photos for visual quality assessment. Specifically, we propose in this paper a Scene-Dependent Aesthetic Model (SDAM) to assess photo quality, by jointly leveraging the geo-context and visual content. Geo-contextual leveraged searching is performed to obtain relevant images with similar content to discover the scene-dependent photography principles for accurate photo quality assessment. To overcome the problem that in many cases the number of the contextually searched images is insufficient for learning the SDAM, we adopt transfer learning to utilize auxiliary photos within the same scene category from other locations for learning photography rules. Extensive experiments shows that the proposed SDAM scheme indeed improves the photo quality assessment accuracy via leveraging photo geo-contexts, compared with traditional universal aesthetic models.

Index Terms— Photo quality assessment, geo-context, transfer learning, social media.

1. INTRODUCTION

Automatic photo quality assessment has drawn numerous research attention in recent decades due to its potential need in various applications. In many media applications, it is desired to single out high quality images. For example, in image retrieval scenario, it is desired for the search engine to be capable of retrieving the images not only by content relevance but also by image quality level. In the social media websites such as Flickr, it is useful to recommend the newly uploaded high quality photo to more users. Therefore, there is a pressing need to automatically assess the quality of images.

Most existing works on image quality assessment aim to learn a universal aesthetic model (UAM) utilizing massive images containing significantly different content. As shown in the scheme 1 of Figure 1 (a), under the assumption that with some universal photography principles, given any input image, the image quality can be assessed by applying the learned UAM, regardless of what content the picture contain. Hence, based on various hand-crafted features which would correlate with several common known photography rules, they either predict image aesthetic score by regression or discriminate high quality photos and low quality ones by classification approaches. In

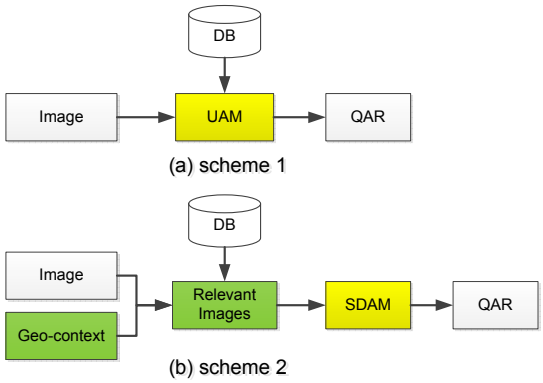

Fig. 1. Two schemes for image quality assessment. (a) Scheme 1: universal aesthetic model (UAM) without context, (b) Scheme 2: our proposed scene-dependent aesthetic model (SDAM) with context (DB: Database, QAR: Quality Assessment Result).

[1], aesthetically pleasing and displeasing images are discriminated using low level visual features such as intensity, color, texture shape and region composition. In [2], the rating confidence proportional to the number of ratings was accounted into the weighted learning procedure to improve the photo quality prediction performance using the same set of low level features as in [1]. In [3], by determining perceptual factors that distinguish professional photos and snapshots, high level features such as spatial distribution of edges, color distribution, and blur were adopted for photo quality classification. Considering that professional photographers often skillfully differentiate the main subject of the photo from the background, blur detection and features such as clarity contrast,lighting difference and subject composition were adopted in [4] for quality classification. In particular, the rule of thirds was adopted in [5] to assess and enhance photo quality based on subject composition. Visual attention model based on saliency map is deployed for photo assessment in [6]. In [7], high level describable image attributes including compositional attributes, content attributes and sky illumination attributes are designed for photo quality prediction.

However, in real cases when professional photographer take pictures, instead of apply several simple photography rules slavishly, they adaptively apply their photography knowledge according to the shooting content. For example, although the rule of thirds is widely used to generate visual features for photo quality assessment, it is not always the optimal composition to place the object on the intersection of the third lines when shooting different objects. As some example high quality photos from Flickr shown in Figure 2, in (a), albeit the rule of thirds works well to determine the position of the Golden Gate Bridge, the symmetric composition which put the object in the middle fit better for the Statue of Liberty. The lighthouse

Fig. 2. Example high quality photos following different photography rules: (a) difference of the object position in the frame, (b) difference of the horizon placement.

is placed far from the image center. Another commonly used photography rule is to place the horizon lower according to the golden ratio to create a sense of stability. However, the reflection photography technique which put the horizon in the middle is also frequently applied as the second picture in (b). Moreover, it encourages a even higher placement of the horizon when the ground has plenty of interest as the third picture in (b). Therefore, photography is truly a scene-dependent process. The direct use of several simple known photography rules are far from sufficient to model the photo aesthetics for their quality assessment. The scene-dependent characteristic of photography make it desirable to build a scene-dependent aesthetic model for accurate photo quality assessment.

1.1. Challenges and Contributions

However, assessing photo quality by building scene-dependent aesthetic models is a great challenge. First, photographers are really artists whose knowledge is difficult to extract and represent with simple rules. The non-exhaustiveness of photography rules make it impossible to model the photo aesthetics based on a list of handcrafted features directly correlated with the limited number of rules. Second, different type of photography principles should be applied to different scenes. Third, even within the same scene category, various photography rules apply when capturing objects with various arrangements and shapes. Professional photographers take years of training to obtain the photography knowledge and they usually carefully adapt the knowledge to capture the scene under different conditions.

However, the context information and social data associated with the photos in the social media community bring up new opportunity for solving this challenging problem. Photos containing the same scene from the same location are very likely to follow similar photography rules. Motivated by this, we propose a paradigm shifting scene-dependent aesthetic model (SDAM) based photo quality assessment scheme by leveraging the photo geo-contexts, which address the challenge from a totally new perspective, compared with the traditional UAM based scheme. As illustrated in Figure 1 (b), by jointly utilizing the input image content and associated geo-context, relevant online images containing similar content are retrieved to learn the SDAM to assess the input photo quality.

The main contributions of this research are summarized as follows. First, we propose a SDAM based photo quality assessment scheme by jointly utilizing the image content and geo-context. To the best of our knowledge, this is the first attempt to leverage geo-context of photographs for photo quality assessment. Second, when inadequate contextual searched relevant images are available for photography rules learning, we adopt transfer learning [8] to effectively utilize the related images in the same scene category from other geo-locations as auxiliary data to improve the photo quality assessment accuracy.

The rest of the paper is organized as follows. Section 2 describes the proposed SDAM based photo quality assessment scheme. Section 3 presents the experiments and evaluations, followed by the conclusions in Section 4.

2. SDAM-BASED PHOTO QUALITY ASSESSMENT

The proposed SDAM based photo quality assessment scheme is based on two observations. We observe that the photos with the same geo-context likely contain same objects within the same scene, since most people tend to take pictures from some photography-worthy viewpoints when visiting a specific destination. Therefore,relevant photography knowledge from these photos can be learned to assess photo quality. We also observe that photography rules applied to photos in the same scene category, but from different locations can also be partially applied to assess photo quality, since photos from different geo-locations but within the same scene category usually follow similar photography rules because of the similar scene structures.

Given a new picture with its geo-location, we aim to assess its quality by learning the SDAM from two types of photos: i)the photos with the same scene from the same location, and ii)the photos with similar scene from other geo-locations when there are not enough relevant photos from the given geo-location. The proposed SDAM based photo quality assessment scheme is illustrated in detail in Figure 3.

To learn the SDAM which correlate with the underlying relevant photography knowledge for the input image quality assessment, we first perform contextual image retrieval to obtain photos containing similar content from the same geo-location. Once sufficient relevant photos in the same geo-context are obtained, conventional machine learning approaches are applied to discover the specific photography rules of the given scene for photo quality assessment. However, except in some truly hot spot locations, the number of photos with the same content in the same geo-context can be quite limited. Such limited samples of photos may pose a great challenge for the learning to be effective. When a photographer takes picture in a new location he has never been before, he usually apply his photography knowledge he has acquired in the similar scene before. Therefore, when relevant contextual searched photos are not sufficient, we need to adaptively transfer the photography knowledge by utilizing photos with similar scene from other locations. Although the photography knowledge learned from photos with similar scene but different geo-contexts is closely related to the desired knowledge, they are not the same. Therefore, transfer learning [8] is adopted to appropriately utilize the related photos from other locations to build the photo quality classifier. Scene recognition is carried out on the input image. Then the photos of the same scene category from other locations are utilized as auxiliary data to learn the SDAM. Finally the input image can be assessed with the learned SDAM which models the underlying relevant photography knowledge.

2.1. SDAM Learning from Contextually Searched Photos

2.1.1. Contextual image retrieval

Images taken from the same location has high probability to contain the same content with the input image. Therefore, relevant pho-

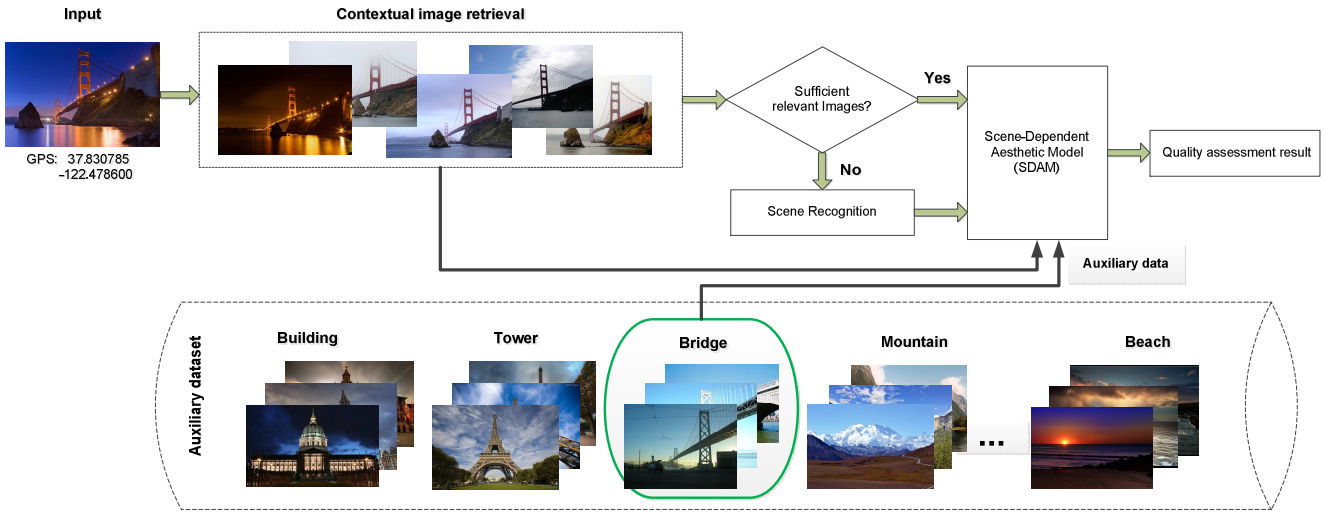

Fig. 3. The proposed SDAM based photo quality assessment scheme.

tography knowledge can be learned from the contextually searched images. Hence, contextual searching is performed to crawl images captured within a certain scope of neighborhood of the input GPS. In the proposed system, we set the radius of search scope as one kilometer in Flickr photo search.

However, images from the same location may contain different scenes. For example, when people are standing near the Golden Gate Bridge, some may shoot the bridge, while some may capture the statue on the seaside. Hence, the contextually searched photos containing different scenes have to be filtered out. We utilize GIST descriptor [9] to further select the images containing the same scene from the contextual searched data, as such descriptor has been widely used and shown to obtain good results in scene recognition. The GIST descriptor computes oriented edge filter responses at different scales aggregated into spatial bins. The scene structures within an image can be characterized by the GIST descriptor. We can measure the similarity between the scene structures presented in two images by comparing their GIST descriptors. We compute the GIST descriptors for the input image and the contextual searched images. The searched images with large L2 distance to that of the input image are filtered out. In our experiment, we removed images with distance greater than the threshold 0.5.

2.1.2. *Image representation*

To learn the relevant photography rules, we attempt to represent images by analyzing their compositions and color distributions, since they are two key factors determining the image aesthetics.

To model the image composition, the arrangement of structures and geometric patterns within an image are represented by GIST descriptor and the spatial layout of saliency. The GIST descriptor measures the oriented edge response at multiple scales, and aggregates the responses into spatial bins to capture the structure features. Besides the scene structure, we also need to model the position of salient objects in the photo, since the salient objects usually draw human visual attention and thus the location and scale of them are also important compositional factors.To model the layout of the salient objects, we compute the saliency map using the spectral residual approach [10] and partition the saliency map into 6×4 spatial bins. As the saliency is sensitive to image scales, we use two scales, 32 and 64, to calculate the saliency map.The sum of the saliency values in each spatial bin for each scale are computed.

In photography and color psychology, color tones, saturation and lighting play important roles. For example, it has been shown that professional photographers tend to use complementary colors and high saturations to make pleasing photos. Also, they usually carefully select the time, direction and camera parameters to set good lighting conditions. Hence we model the color distributions in HSV color space.

2.1.3. *Contextual SDAM learning*

For input images from hot spot scenery locations, the quantity of contextually searched relevant images containing the same scene with the input are usually sufficient. we rank the relevant searched images by their aesthetic scores. In this work, we focus on discriminating high quality and low quality photos, hence the top 10% and the bottom 10% ranked images are treated as positive and negative training data. Going further down in the ranking is possible, but increases ambiguity in photo quality. Regression model can also be built using the relevant images to predict photo aesthetic scores.

Although explicit aesthetic score is not available to photos in flickr, the large volume of social information in the social media website can reflect the photo appealing degrees. From observation, high quality photos usually draw much attention from the community, hence having large number of views and favors. Therefore we calculate the aesthetic score based on the number of views and number of favors on the photos. For image *I*, we approximate its aesthetic score S_I by

$$S_I = 100 \times \left(1 - \exp\{-(\alpha \cdot views_I + \beta \cdot favors_I)\}\right) \quad (1)$$

where $views_I$ and $favors_I$ are the number of views and the number of favors of image I respectively. α and β are two coefficients to weight the number of views and favors. Here, we use $\alpha = 0.2$ and $\beta = 1$ in our experiments. Using the above equation, the scores of images are ranged from 0 to 100. According to the aesthetic score, the contextual searched relevant images are ranked to generate the positive and negative training set.

Once the image representation is complete, conventional machine learning methods can be utilized to build the SDAM. In the

proposed system, we apply support vector machine (SVM) to find the optimal hyperplane based on the extracted features for quality assessment.

2.2. SDAM Learning by Leveraging Auxiliary Data

Although learning photography knowledge from contextually retrieved images works well, in many less well-known geo-locations, only small number of relevant photos can be retrieved. This is not sufficient for learning. However, there exist vast number of images from similar scene but taken from different locations are still available. These images can be utilized to learn the photography rules for a specific scene in the given location, even though the rules may not be exactly the same as the input scene. To utilize relevant auxiliary data from other locations, we introduce transfer learning to build the SDAM for the input scene.

2.2.1. Auxiliary dataset building

In this work we mainly focus on scenery photo quality assessment problem, hence we build a large image dataset containing eight most common scene categories: building, tower, bridge, beach, mountain, forest, waterfall and field by keyword based searching from Flickr. For each scene, we rank the searched images by their aesthetic scores. The top 2,000 and bottom 2,000 ranked photos are taken as positive and negative auxiliary data, respectively. Though the positive data and negative data can be generated by simply splitting the aesthetics based ranking list into two halves, the photos ranked in the middle increase the ambiguity in ratings. These images are also ranked by their aesthetic scores calculated with equation (1).

2.2.2. Scene recognition

To obtain related data, we utilize photos in the same scene category as auxiliary data since photographers usually adopt similar rules in the same scene category. Therefore, scene recognition is first performed on the input image. As GIST descriptor can measure the similarity of two scene structures between two images, we calculate the L2 distance of the GIST descriptor of the input image and the images of the auxiliary dataset. Then K-nearest neighbor (KNN) method is applied to classify the input photo into the scene category it belongs to. K is set to be 10 in our experiment.

2.2.3. SDAM learning using auxiliary data

Due to the inadequate quantity of contextual retrieved data, we wish to adopt the photography knowledge of the same scene category but from other locations for quality assessment of the input photo. Although photographers usually apply their own photography knowledge learned from similar locations to the current scene, they will definitely adjust the photography rules accordingly based on the specifics in the new location, including lighting condition and different arrangement or shape of the objects in the scene. Hence traditional learning methods cannot be directly applied when the quantity of contextually retrieved photos is inadequate. Therefore we apply transfer learning [8] in order to appropriately utilize the related auxiliary data in the same scene category from other locations.

When transferring the photography knowledge into the input scene, we have to carefully select the auxiliary data, since some of them share similar photography rules while others do not, even though they belong to the same scene category. Therefore, we iteratively update the weights of the auxiliary data to transfer proper photography knowledge into the current photo quality classifier by transfer Adaboost learning method.

Algorithm 1 SDAM learning via auxiliary data

Input: training set $T = \{x_i, c(x_i)\}$.

1: Initialize the weight vector $w^1 = (w_1^1, \ldots, w_{n_a+n_c}^1)$ as uniform distribution.
2: **for** $t = 1, \ldots, N$ **do**
3: Set $p^t = w^t / (\sum_{i=1}^{n_a+n_c} w_i^t)$.
4: Call the weighted SVM, providing it the combined training set T with the distribution p^t over T to get the hypothesis h_t.
5: Calculate the error of h_t on X_c.

$$\epsilon_t = \sum_{i=n_a+1}^{n_a+n_c} \frac{w_i^t \cdot |h_t(x_i) - c(x_i)|}{\sum_{i=n_a+1}^{n_a+n_c} w_i^t}$$

6: Set $\beta_t = \epsilon_t/(1-\epsilon_t)$ and $\beta = 1/(1+\sqrt{2\ln n_a/N})$. ϵ_t is required to be less than $1/2$.
7: Update the weight vector

$$w_i^{t+1} = \begin{cases} w_i^t \beta^{|h_t(x_i)-c(x_i)|}, & 1 \leqslant i \leqslant n_a \\ w_i^t \beta_t^{-|h_t(x_i)-c(x_i)|}, & n_a+1 \leqslant i \leqslant n_a+n_c \end{cases}$$

8: **end for**
Output:

$$h(x) = \begin{cases} 1, & \texttt{if } \prod_{t=\lceil N/2 \rceil}^{N} \beta_t^{-h_t(x)} \geqslant \prod_{t=\lceil N/2 \rceil}^{N} \beta_t^{-\frac{1}{2}} \\ 0, & \texttt{otherwise}. \end{cases}$$

Let X_c denote the n_c contextually retrieved photos that share the same photography rules with the input image, X_a denote the n_a auxiliary photos in the same scene category from other geo-locations, some of the auxiliary photos may share similar photography rules as the input. The contextually retrieved data and the auxiliary data from other locations are combined to build the training set $T = \{x_i, c(x_i)\}$.

$$x_i = \begin{cases} x_i^a, & i = 1, \ldots, n_a \\ x_i^c, & i = n_a+1, \ldots, n_a+n_c \end{cases} \quad (2)$$

where c is the Boolean mapping function, in which $c(x) = 1$ and $c(x) = 0$ indicate high quality and low quality respectively. The auxiliary data leveraged photography learning algorithm is described in Algorithm 1.At each iteration of the boosting process, instance selection is carried out by adjusting the weight to filter out the images following different photography rules. In particular, the incorrectly classified contextually searched image are weighted higher to draw more attention in the next round, while the incorrectly classified auxiliary data from other geo-locations are weighted lower, since these auxiliary photos could be those following different photography rules from the input scene. Hence, the weight of the mistakenly predicted auxiliary data is decreased through multiplying by $\beta^{|h_t(x_i)-c(x_i)|}$. Thus the auxiliary photos following different rules will have less effect on the learning process in the next iteration.

3. EXPERIMENTS

To verify the proposed scheme, we carried out experiments on a dataset of around 9,600 geo-tagged photos containing a variety of

scenes. The auxiliary dataset is built by top 2,000 and bottom 2,000 ranked photos for each scene category. To build the dataset, we performed geo-location based searching from 16 geo-locations with radius of one kilometer. Eight locations are hot spot places, from where 3,000–8,000 photos were obtained for each location, while the others are sparse locations, where only 30–400 photos were obtained for each location. Most photos from these locations belong to one of the eight defined scene categories, while a few photos contain some unique content such as some buildings or statues of special shapes. As described before, the top 10% and the bottom 10% ranked contextual searched relevant images are utilized for SDAM learning for a given input image. Hence, for each location, the top 10% and the bottom 10% ranked images are treated as high quality and low quality photos, respectively. In the dataset, about 9,000 are from hot spot locations and about 600 photos are from sparse locations.

In the experiment, each image with its associated geo-context is fed into the developed SDAM based photo quality assessment system as the input, and then relevant images with similar geo-contexts are obtained to learn SDAM for image quality assessment. When the quantity of relevant images are not sufficient, auxiliary data of the corresponding scene is utilized for SDAM learning. In our experiment, once the number of the relevant images are smaller than 500, the auxiliary data are used. To compare the performance of the proposed SDAM-based scheme and traditional UAM-based scheme, we also implemented UAM based photo quality assessment scheme using the same set of features to make the comparison reasonable. Since UAM-based photo assessment scheme learns the universal aesthetic model regardless of the image scene categories and image geo-contexts, auxiliary dataset of 32,000 images belonging to the eight different scene categories are utilized as the training set.

The accuracy comparison of UAM scheme and the proposed SDAM scheme is listed in Table 1 . The traditional UAM-based assessment scheme only achieved 67% and 64% accuracy on images from hot spot locations and images from sparse locations, respectively. The proposed SDAM scheme achieved 81% and and 73% accuracy for images from hot spot locations and images from sparse locations. The comparison results show that the SDAM-based photo assessment scheme indeed outperforms the traditional UAM-based scheme by leveraging the photo geo-context and auxiliary images of the same scene categories. The UAM-based scheme works a little bit better on images of hot spot locations than images of sparse locations, probably due to the existence of some training images with similar content from the same locations. In addition, we also performed the SDAM-based assessment on images of sparse locations without using the auxiliary data. Due to lack of sufficient relevant images for SDAM learning, the accuracy dramatically decreased to 57%, compared with the auxiliary data leveraged accuracy of 73%. Therefore, although the performance of SDAM learning using auxiliary data when contextual searched relevant data is limited, is not as good as the conventional learning approach when relevant training data is sufficient, the adoption of the auxiliary data from other places indeed helps to overcome the insufficient training data problem and greatly improves the quality assessment results.

Due to the page limit, we only demonstrate some examples of UAM and SDAM based assessment results on the highest and lowest ranked three photos from four hot spot locations and four sparse locations in Figure 4. In each group, the first three and last three rows are top and bottom ranked three images, respectively. The left column is the SDAM-based assessment results, in which the misclassified photos are highlighted in red rectangle; while the the right column is the UAM-based assessment results, in which the misclassified photos are highlighted in yellow rectangle. Without learning input scene specific photography rules, UAM can only model some general composition and color principles for photography, which may not suitable for all images with various scenes under different lightings. Therefore, more images are assessed incorrectly. Compared with the UAM based scheme, the proposed SDAM based assessment scheme learns more relevant photography rules for input image, hence the assessment results are greatly improved.

Table 1. The accuracy comparison of UAM and proposed SDAM based photo quality assessment schemes.

Accuracy	Hot-spot locations	Sparse locations
UAM	67%	64%
SDAM	81%	73%

4. CONCLUSION

In this paper, we proposed a SDAM based photo quality assessment scheme by leveraging context information. From experiments, the leveraging of context indeed helps to learn input scene relevant photography rules for better quality assessment, compared existing traditional UAM based assessment scheme. The adoption of auxiliary data in the same scene category indeed benefits the SDAM learning when relevant contextual searched data is not sufficient. In the future, regression model can also be applied in the SDAM to improve the aesthetic score prediction performance. Also, many other visual features such as SIFT features or face detection can be utilized in the scene dependent aesthetic model to extract various underlying photography principles for aesthetic quality classification or regression.

5. REFERENCES

[1] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Ze Wang, "Studying aesthetics in photographic images using a computational approach," in *Proc. of European Conference on Computer Vision*, 2006, pp. 288–301.

[2] Ritendra Datta, Jia Li, and James Ze Wang, "Learning the consensus on visual quality for next-generation image management," in *Proc. of ACM Multimedia*, 2007, pp. 533–536.

[3] Yan Ke, Xiaoou Tang, and Feng Jing, "The design of high-level features for photo quality assessment," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 419–426.

[4] Yiwen Luo and Xiaoou Tang, "Photo and video quality evaluation: Focusing on the subject," in *Proc. of European Conference on Computer Vision*, 2008, pp. 386–399.

[5] Subhabrata Bhattacharya, Rahul Sukthankar, and Mubarak Shah, "A framework for photo-quality assessment and enhancement based on visual aesthetics," in *Proc. of ACM Multimedia*, 2010, pp. 271–280.

[6] Xiaoshuai Sun, Hongxun Yao, Rongrong Ji, and Shaohui Liu, "Photo assessment based on computational visual attention model," in *Proc. of ACM Multimedia*, 2009, pp. 541–544.

[7] Sagnik Dhar, Vicente Ordonez, and Tamara L. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1657–1664.

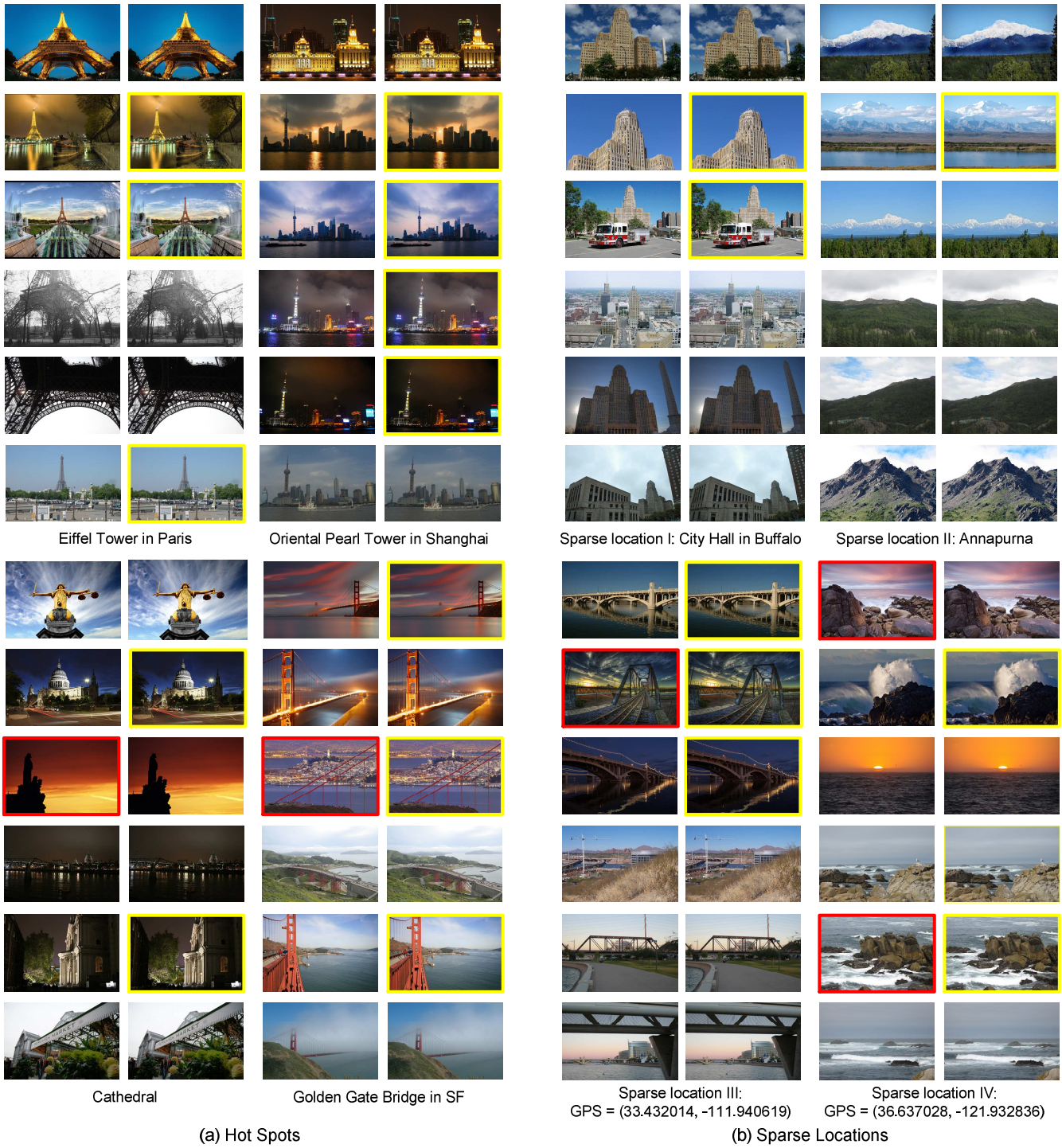

Fig. 4. Examples of quality assessment results by UAM and the proposed SDAM schemes, from the photos from (a) hot spots and (b) sparse locations. The first three and the last three rows are top three (with high visual quality) and bottom three (with low visual quality) ranked photos, respectively. The incorrect results from SDAM and UAM are highlighted in red and yellow rectangles, respectively.

[8] Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu, "Boosting for transfer learning," in *Proc. of International Conference on Machine Learning*, 2007, pp. 193–200.

[9] Aude Oliva and Antonio Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[10] Xiaodi Hou and Liqing Zhang, "Saliency detection: A spectral residual approach," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2007.