Viewpoint Recommendation Based on Object-Oriented 3D Scene Reconstruction

Ke Li, Yuxia Wu, Yao Xue, and Xueming Qian, *Member, IEEE*

***Abstract*—Viewpoint recommendation can recommend several viewpoints for taking aesthetic photographs of a place-of-interest (POI) and is of great importance for photography assistance. In this paper, we propose a system that can assist a user in choosing good viewpoints for taking high-quality photographs. Our system is based on social media and 3D reconstruction. To reduce the time cost and improve the quality of 3D reconstruction, we propose a weakly supervised object detection method that is used before 3D reconstruction. The camera pose of images is recovered by the subsequent 3D reconstruction pipeline. We use a convolutional neural network (CNN) to extract 2D image features, and we fuse them with 3D camera pose features to learn their relationships to image aesthetics. The trained model is utilized to evaluate the aesthetics of images. Finally, the 3D space of all possible camera poses is divided into 3D grids, and the aesthetics score of each grid is evaluated. We combine the aesthetics and diversity of all viewpoints and recommend several high-quality viewpoints. Experimental results indicate that our approach can help users choose viewpoints that will result in high-quality photographs while maintaining diversity.**

***Index Terms*—3D reconstruction, social media, aesthetics evaluation, viewpoint recommendation.**

I. Introduction

With the development of social media networks, an increasing number of photos are available on the Internet [1]. The number of images for every place-of-interest (POI) is growing at an amazing speed, which gives us crowdsourced data. When people are traveling around the world, they are likely to take many photos and share them with others on websites such as Flickr and photo.net. People may comment on photos uploaded by others, which raises the following question: which photos have enough aesthetic quality to receive positive comments? Different people may have various answers to this question, but one factor must play a central role: the viewpoint. Choosing a good viewpoint is a difficult job for photographers, especially for new users and for a POI that a user is not familiar with. It would be helpful if there were a method that could recommend a viewpoint for taking photos; thus, there is a need for viewpoint recommendation.

Some rules have been summarized for ordinary users to take good photos [2], but experienced users have professional knowledge. To estimate the aesthetics of images, some models that use crowdsourced social media data have been proposed [3]. Salient object detection is performed in other models [4] to be used by subsequent aesthetic assessment. The problem of salience detection-based methods is that their performance is limited by the quality of the salience map.

Viewpoint recommendation methods that depend on text-based image retrieval usually cannot recommend a good viewpoint. Research communities have proposed several methods to estimate the aesthetics of images [5]. However, those methods focus on image content, composition or high-level features, and they are not specially designed for viewpoint recommendation.

Image aesthetics is an important aspect of viewpoint recommendation. Many factors contribute to the aesthetics of a photograph, such as illumination, composition, weather conditions and viewpoint. The research community has studied methods to estimate image aesthetics [4], [6] and measure the composition of photos. [6] proposes an intelligent photography system that can function as an intelligent professional view guide based on real-time view quality assessment and an embedded compass. Their proposed system mainly focuses on professional photo composition. However, these methods consider only 2D image features. The 3D position from which an image is taken can also affect image aesthetics. These methods cannot reflect the importance of the viewpoint for taking photos.

To solve the above problems, we propose a new viewpoint recommendation method that utilizes weakly supervised object detection-based 3D reconstruction from crowdsourced social media. Our system results in more visually attractive images by considering both aesthetics and diversity. It is inspired by the following ideas: 1) Different people have different preferences for the viewpoint of a POI, so both aesthetics and diversity should be considered. 2) There are plenty of images of various quality on the Internet, which makes it difficult and necessary to mine aesthetic images from crowdsourced data. 3) 3D reconstruction can recover the camera pose of photos and thus may play a central role in viewpoint recommendation.

Manuscript received November 28, 2019; revised February 7, 2020 and March 10, 2020; accepted March 10, 2020. Date of publication March 18, 2020; date of current version December 17, 2020. This work was supported in part by the NSFC under Grants 61732008 and 61772407, in part by the Guangdong Provincial Science and Technology Plan under Grant 2016A010101005, and in part by Microsoft Research Asia. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sebastian Knorr. *(Corresponding author: Xueming Qian.)*

Ke Li is with the School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: likely@stu.xjtu.edu.cn).

Yuxia Wu and Yao Xue are with the School of Information and Communication Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: wuyuxia@stu.xjtu.edu.cn; yxue2@ualberta.ca).

Xueming Qian is with the Key Laboratory for Intelligent Networks and Network Security, Ministry of Education, School of Information and Communication Engineering, and the Smiles Laboratory, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: qianxm@mail.xjtu.edu.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2020.2981237

With the above ideas, we now give a simple introduction to our viewpoint recommendation method. First, we perform weakly supervised object detection before 3D reconstruction, which can save considerable time in 3D reconstruction and improve the 3D reconstruction results. Second, we recover the camera pose of each image in the dataset by structure-from-motion (SfM) [4], [7]. The viewpoint, i.e., the position where an image is taken, is encoded in the camera pose of that image. Third, we design an aesthetic assessment system based on a deep neural network called ViewNet. The relationship between image aesthetics and image features is learned by our model. We use a trained model to calculate the image aesthetics score of each image that is used for 3D reconstruction. Finally, we split the 3D space that contains all possible positions of the camera into 3D grids. We calculate the mean aesthetics of each grid and recommend images from different grids; in each grid, we recommend highly aesthetic images in that grid; thus, both the quality and diversity of viewpoints are ensured.

The main contributions of this paper are as follows:

1) We propose a new viewpoint recommendation method for photography assistance that uses crowdsourced social media on the Internet uploaded by users. We divide the potential space for taking photos into 3D grids. We assess the aesthetic quality of each grid and rank the grids by aesthetics. We only select images from grids with high aesthetics, which ensures the aesthetics of the recommendation results; the images are selected from different grids with different styles, which ensures the diversity of the recommendation results.

2) We perform main object detection before 3D reconstruction, thus saving considerable reconstruction time and achieving better reconstruction results. Our object detection method is weakly supervised. We use a pretrained model to extract features and select the top N channels with the strongest feature map response. Then, the bounding box of the object is determined in the selected channels of the feature map.

3) An image aesthetics assessment model called ViewNet is designed to evaluate the quality of the images that participate in 3D reconstruction. This model jointly learns 3D camera pose features and 2D image content features. The recovered camera pose during 3D reconstruction serves as the basis of our viewpoint recommendation system.

The rest of this paper is organized as follows: Section II gives a brief overview of related work. The implementation of the proposed viewpoint recommendation model is introduced in Section III. Section IV shows the details of our dataset and experimental results. Section V concludes with a summary of our method.

II. Related Work

In this section, we briefly overview the related work on weakly supervised object detection, 3D reconstruction and POI summarization.

A. Weakly Supervised Object Detection

Object detection, i.e., estimating the class and location of objects in an image, has attracted great attention in the research community. Fully supervised approaches have greatly improved the accuracy of object detection [8]–[11]. However, fully supervised object detection methods require detailed annotations, i.e., bounding boxes. It is time-consuming to annotate bounding boxes in large image datasets. This fact leads researchers to consider the weakly supervised object detection (WSOD) problem [12]–[16].

Chum *et al.* [17] proposed an exemplar model that can learn and generate a region of interest around class instances, given only a set of images containing the visual class. Saleh *et al.* [12] proposed an approach to solve the semantic segmentation problem by utilizing networks pretrained for the task of image classification. They first use the features obtained from the higher-level convolutional layers of a network to generate foreground/background masks. Then, they fuse the foreground/background masks with information generated by a weakly supervised localization network to compose multiclass masks. Tang *et al.* [13] proposed a three-stage weakly supervised object detection approach. In the first stage, the method of [13] generates coarse proposals from a dense set of sliding window boxes. Then, [13] refines the generated proposals in the proposal refinement stage to obtain more precise proposals. Finally, the WSOD stage in [13] classifies the refined proposals to generate detection results. [13] utilizes the alternating network training strategy in the Faster RCNN [18], which enables the model to share the weights of conv layers among different stages.

The differences between our method and the above WSOD methods are as follows: 1) Our method selects channels from the feature map with the top N strongest responses, while [12] uses all channels from the fourth and fifth layers of the VGG network. 2) Our method does not require any training, so it only contains a forward-pass stage, while [13] is a three-stage weakly supervised object detection method that utilizes the alternating network training strategy in the Faster RCNN [18]. 3) [12] can obtain accurate class-specific masks for different classes, while our method only obtains a bounding box for foreground objects and is not class-specific. This is reasonable because 3D reconstruction is a later step in our method, so detecting foreground objects and feeding cropped images into 3D reconstruction pipelines is sufficient.

Although object detection approaches have been successfully applied in many computer vision tasks, this is the first time they have been used in 3D reconstruction.

B. 3D Reconstruction

The SfM problem in computer vision is the problem of recovering the three-dimensional (3D) structure of a stationary scene and camera pose from a set of two-dimensional (2D) images.

For sparse reconstruction, SfM [7], [19]–[21] can usually be employed to obtain camera poses and sparse 3D point clouds. For dense reconstruction, MVS algorithms [22] are utilized to generate a dense-patch 3D model.

SfM is the main method for 3D reconstruction. In 2006, Snavely *et al.* [7] proposed a sequential pipeline for SfM, showing that this system can result in high-quality 3D scenes and camera poses in hundreds of unordered images. Zhou *et al.* [19] proposed a shape deformation model to solve the nonrigid SfM problem.

Schonberger *et al.* [21] proposed a new SfM system that introduces a geometric verification method. This method augments the scene graph and improves the robustness of the initialization and triangulation sub-process. A next-best-image selection strategy is also used in [21], which maximizes the robustness and accuracy of the incremental reconstruction process.

Our method is different from the above methods in that we perform weakly supervised main object detection before 3D reconstruction and use cropped small images to reconstruct 3D models. There are two reasons for removing these background points: 1) different images have different backgrounds, which causes more error in feature-matching during 3D reconstruction, and we can obtain better 3D reconstruction results if these background points are removed; 2) matching these background keypoints requires considerable computing time, and these background points provide no additional useful information for 3D reconstruction.

C. POI Summarization

POI summarization aims at recommending several images for a POI. Some methods mainly depend on geographical clustering of images by longitude and latitude so that images with short geographical distances are more likely to be separated into the same cluster [23].

There are already some clustering-based methods. Simon *et al.* [24] proposed a scene summarization method that uses image collections from the Internet. Jiang *et al.* [25] proposed an author-topic model-based collaborative filtering method that is used to make recommendations for social media users. The method recommends images taken from high frequency shooting locations. Qian *et al.* [26] proposed a user-based event summarization method that makes use of different types of data: user, text, and image. They utilize a coarse-to-fine filtering method to eliminate irrelevant information that reduces the dataset's effect. Qian *et al.* [27] proposed a clustering-based method that uses location, appearance, semantic, and temporal information to discover the representative viewpoints of a POI.

Viewpoint recommendation is the problem of choosing good viewpoints for taking photographs [28]. Some works have been done in this field. He *et al.* [28] proposed a robust algorithm that is dedicated to architecture viewpoint recommendation. They designed their system by jointly learning 2D image features and 3D geometric features from images on the Internet. The SVM2K multiview learner is utilized to learn from extracted 2D and 3D features. Rawat *et al.* [29] designed a viewpoint recommendation system called ClickSmart, which is used to help photographers take photos with high aesthetics. ClickSmart uses contextual information such as weather conditions to improve the performance of the viewpoint recommendation system.

Image aesthetics evaluation is the task of analyzing the relationships between the composition of a photo and its aesthetics [23]. Lok *et al.* [30] proposed a bitmap representation of the visual weight of an image, which is called WeightMap. The components in the image are assigned a visual weight, and a WeightMap is designed for encoding the visual weight of the objects in an image.

In contrast to existing methods that require hand-crafted features or only consider 2D image features, our viewpoint recommendation approach uses a CNN to extract image features so that no hand-crafted image features are involved. We propose ViewNet, a deep neural network that jointly learns from 2D image features and 3D pose features. We divide the 3D space into grids and recommend images from grids with high aesthetics scores to ensure aesthetics. We recommend images from different grids, which represent different viewpoints; this ensures diversity.

III. Our Viewpoint Recommendation Approach

We will give an overview of our system and then introduce our method in detail. Our approach mainly consists of the following four steps: 1) weakly supervised object detection, 2) 3D reconstruction, 3) ViewNet training and testing, and 4) viewpoint recommendation. An overview of our system is shown in Fig. 1. The POI dataset is crawled from Flickr. We filter out those images whose GPS locations are too far from the real location of the POI to which they belong. First, we propose a weakly supervised method to detect the main object in each image. In our method, no bounding box annotations are needed, and we use a pretrained VGG16 model (we remove the final fully connected layers to be able to take random-sized images as input without resizing the input images) to extract features from the whole dataset. We analyze the output of the max-pooling layer of all images and select the top N channels with the strongest response for detecting the bounding box. Then, we crop the foreground object from images and feed the cropped images into a 3D reconstruction system to obtain a 3D model of the POI. Third, we design an aesthetic evaluation model called ViewNet to assess the aesthetics of POI images. ViewNet can jointly learn from 2D image features and 3D camera pose features. The 2D image features are extracted by a CNN, and 3D camera pose features are obtained by 3D reconstruction. Finally, we divide the 3D space that contains all camera positions into grids. We evaluate the aesthetics of the images in each grid and calculate the mean aesthetics of each grid. We consider aesthetics and diversity for viewpoint recommendation. We evaluate the aesthetics of each grid by the previously trained model, and we achieve diverse image summarization results by recommending photos from different grids.

A. Weakly Supervised Main Object Detection

We define N_{fo} as the number of feature points detected in the main object of an image, such as an architectural structure, and N_{fi} as the number of feature points detected in the whole image. Then, the ratio of object feature points, R_{oi}, can be derived as Equation (1):

$$R_{oi} = N_{fo}/N_{fi} \tag{1}$$

We calculate the R_{oi} of every image in each POI, and then we obtain the mean R_{oi} of each POI, which is shown in Table I. The results in Table I mean that the POIs have many keypoints that are detected in the background area, and 3D reconstruction speed and quality will be improved if we remove these background

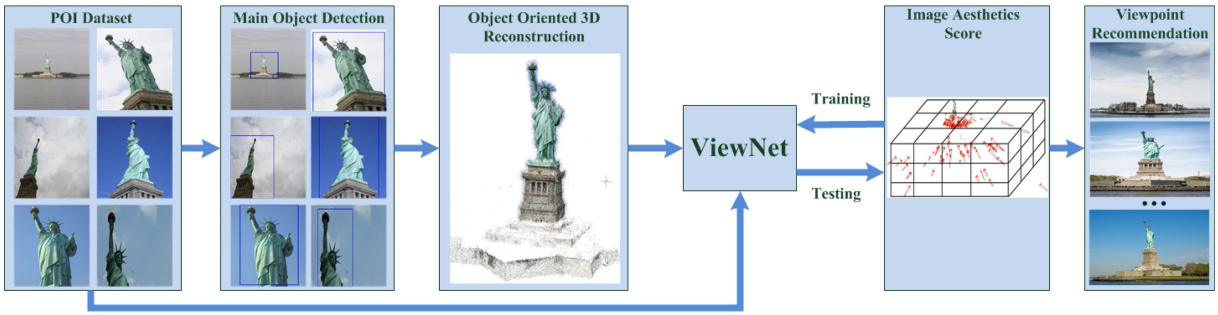


Fig. 1. System overview of our viewpoint recommendation method based on 3D scene reconstruction.The detailed structure of ViewNet is shown in Fig. 6.

TABLE I
MEAN R_{oi} OF DIFFERENT POIS

POI	Mean R_{oi}
#1 Arc de Triomphe	0.74
#2 Big Ben	0.65
#3 Cologne Cathedral	0.85
#4 Eiffel Tower	0.72
#5 Leaning Tower of Pisa	0.64
#6 Mount Rushmore	0.49
#7 Statue of Liberty	0.73
#8 Taj Mahal	0.62
#9 Tiananmen	0.56

points. This shows the necessity for main object detection before the SfM step.

3D reconstruction is a time-consuming procedure. The existence of background keypoints can cause more errors in matching because, under most conditions, the backgrounds of different images in the same POI are different [31]. In addition, matching these background keypoints increases the time cost of 3D reconstruction.

To solve these two problems, we introduce a weakly supervised object detection procedure before performing SfM. Our weakly supervised object detection method consists of the following steps: 1) select the top N channels from the feature map [35] and 2) generate bounding boxes. We now discuss these steps in detail.

1) Select the Top N Channels From the Feature Map: We take a pretrained VGG16 as our feature extractor and remove the fully connected layers to make the model able to take random-sized images as input without resizing the input images. Each image will obtain a feature map F, where F is the output of the last max-pooling layer and it is a 3D tensor (the three dimensions are channel, width and height). We calculate the sum of all the channels of F in every image with Equation (2):

$$FS = \sum_{j=1}^{H} \sum_{k=1}^{W} F_{ijk} \tag{2}$$

where H is the height of F and W is the width of F. We call FS the *sum feature vector* of an image; every image will have an FS, and its dimension will be equal to the number of channels in F. Then, we calculate the mean of all FS features in the channel direction:

$$FM = \frac{\sum_{i=1}^{M} FS_i}{M} \tag{3}$$

where M is the number of images in our dataset, FS_i is the *sum feature vector* of the i-th image, and the operation in the numerator of Equation (3) is vector addition. We call FM the *mean feature vector* of the dataset, and its dimension is also equal to the number of channels in F.

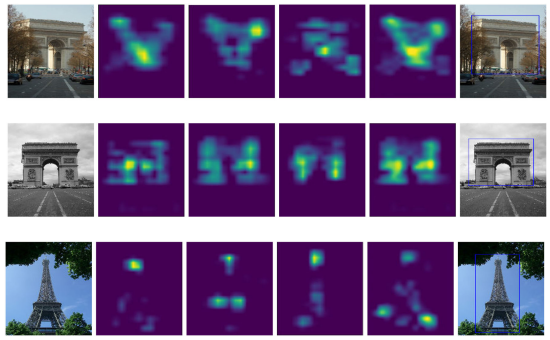
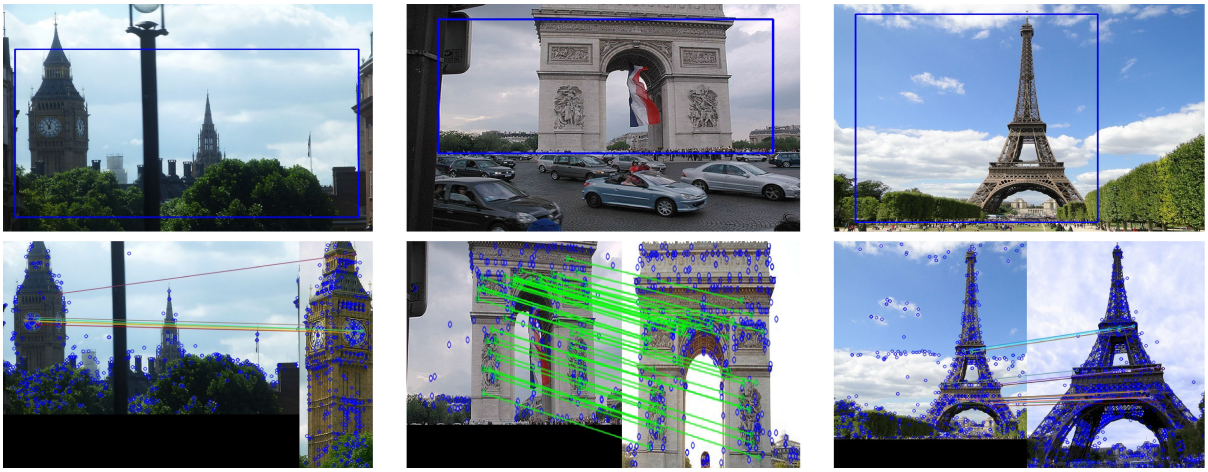


Fig. 2. Visualization of selected top N channel responses. Columns 2–4: the top 3 channels. Column 5: mean value of the top N channels. Column 6: detection result.

We select the top N channels, where FM has the top N largest values. The reason is that these channels carry more information than the other channels. Fig. 2 shows the visualization of the top N channels of the feature map.

2) Generate Bounding Boxes: For each image's feature map F, we select the top N channels according the above method and obtain a mean feature map of the top N channels. We denote the mean feature map as F_d, which will be used for detecting foreground objects in the original image. Then, we resize F_d to the size of the input image and denote the resized feature map as F_D. We take the average value of all elements in F_D as a threshold and denote the threshold by θ. We scan F_D in the x direction and y direction to find elements larger than θ. In the x direction, we record the x coordinates of the first and last elements that are larger than θ, and we denote them by $Xmin$ and $Xmax$; in the y direction, we do the same thing and obtain $Ymin$ and $Ymax$. Thus, $(Xmin, Ymin)$ and $(Xmax, Ymax)$ determine a bounding box in the original image. The last column of Fig. 2 shows some results of our weakly supervised object detection approach.



(a) Big Ben (b) Arc de Triomphe (c) Eiffel Tower

Fig. 3. The first row shows some failure cases in main object detection. The second row shows matching results of failure cases and success cases.

3) Discussion of the Failure Cases of Main Object Detection: As our main object detection is weakly supervised, there are some failure cases in our detection results, which are shown in the first row of Fig. 3. The detection procedure is the preprocessing of 3D reconstruction rather than our ultimate aim. Failure cases in the detection procedure are acceptable in the later 3D construction pipeline. There are two reasons for this: 1) if the detected bounding box of an image is not accurate and the cropped image contains only parts of the foreground, other correctly cropped images can compensate for the lost information; 2) if the cropped image contains not only the foreground but also part of the background, the feature extraction algorithm will extract some keypoints that correspond to the background object. Feature matching is a necessary stage in the 3D reconstruction pipeline. During the feature-matching stage, there are insufficient background keypoints in other images to be matched with these background keypoints. Therefore, the effect of incorrectly detected bounding boxes can be eliminated during the feature-matching stage. This is illustrated in the second row of Fig. 3.

B. Object-Oriented 3D Reconstruction

We crop the main object from the original images with our weakly supervised object detection method. We use the cropped images as the input of the SfM algorithm. In our paper, we use COLMAP [21] as our SfM pipeline. COLMAP uses RootSIFT [32] to extract features and works well for datasets crawled from the web. Fig. 4 shows the results of SfM for the Statue of Liberty POI. Fig. 5 shows several dense 3D reconstruction results of some world-famous POIs.

C. ViewNet Training and Testing

As shown in Fig. 4, it is obvious that people have different preferences when taking photos for a POI. Some viewpoints are selected by more people–thus, there are more cameras in the reconstructed model–while some other viewpoints have fewer cameras. It is reasonable to assume that this phenomenon results from the different aesthetics of different viewpoints.

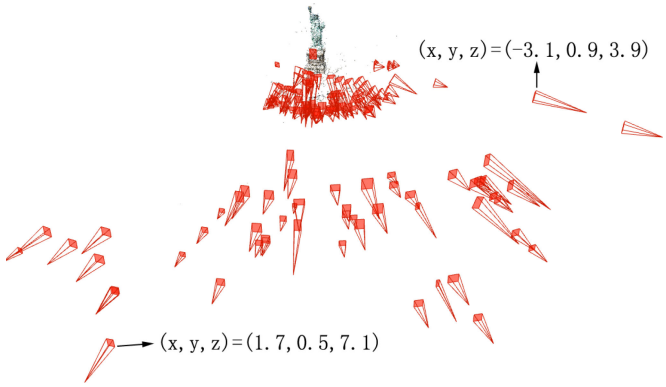
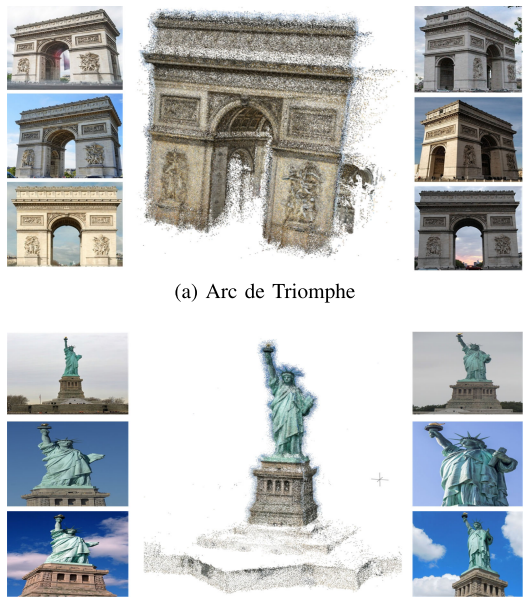


Fig. 4. Result of sparse 3D reconstruction, each pyramid represents a camera.



(a) Arc de Triomphe

(b) Statue of Liberty

Fig. 5. 3D reconstruction results of some example POIs. The reconstructed 3D model is surrounded by images.

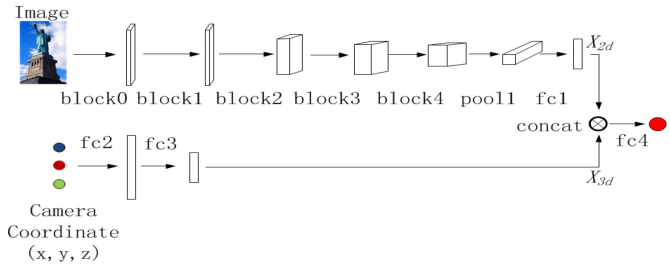


Fig. 6. Network structure of ViewNet, its output is a scalar.

TABLE II
DETAILED CONFIGURATION OF VIEWNET

name	detail
block0	[7x7 conv, 64, /2], [3x3 maxpool, /2]
block1	[3x3 conv, 64], [3x3 conv, 64]×3
block2	[3x3 conv, 128, /2], [3x3 conv, 128]×3
block3	[3x3 conv, 256, /2], [3x3 conv, 256]×3
block4	[3x3 conv, 512, /2], [3x3 conv, 512]×3
pool1	[7x7 avgpool]
fc1	[fc, 10]
fc2	[fc, 15]
fc3	[fc, 10]
fc4	[fc, 1]

1) Network Architecture: To model the above phenomenon, we propose ViewNet, a deep neural network that is capable of learning to assess and recommend image aesthetics with camera pose embedding. It is shown in Fig. 6.

The network consists of two branches: a) a 2D image feature branch and b) a camera pose embedding branch.

a) The 2D image feature branch is composed of different arrangements of CNNs, batch normalization layers, activation layers, max-pooling layers and fully connected layers. The input images are the original, uncropped images corresponding to the cropped images that were previously used for 3D reconstruction. The input images are resized to 224×224 to serve as the initial input of the 2D image feature branch. Table II shows the detailed configuration of this branch.

b) The camera pose embedding branch consists of the following layers: a fully connected layer that has an input of size 3 and output of size 15, a ReLU activation layer and a fully connected layer that has an input of size 15. Its output size is the same as that of the 2D image feature branch. This branch is simple, and Table II shows the detailed configuration of this branch.

Then, the features from two branches are concatenated to obtain a longer vector that contains both 2D features and 3D features. Finally, the concatenated vector is fed into another fully connected layer and activated by a sigmoid function to obtain the aesthetics score.

We record the mean value of the 2D branch output and 3D branch output, as well as the standard deviation of the 2D branch output and the 3D branch output, during the training stage. The output of the 2D image branch is in the range of 0.38 ± 0.22, and the output of the 3D camera-embedding branch is in the range of 0.5 ± 0.09. We can infer that the difference in output between the 2D branch and 3D branch is not very large. This means that the 3 * 1 input is not ignored, because its output range is similar to that of the 2D input branch.

2) ViewNet Training: ViewNet treats the image aesthetics assessment problem as a regression problem. The output of ViewNet is activated by a sigmoid function, so the output is in the interval [0, 1]. We need to shift the output to the interval [1, 4] to be compatible with the ground truth:

$$Z = 3 * Y + 1 \tag{4}$$

where Y is the output of ViewNet and Z is the final predicted score. As a regression problem, we use the mean-squared error as the cost function:

$$J = \frac{1}{N}\sum_{i=1}^{N}(Z_i - \hat{Z}_i)^2 \tag{5}$$

where Z_i is the ground truth score, $\hat{Z}_i$ is the predicted score, and N is the batch size; we set $N = 16$ in our experiment. We use the stochastic gradient descent algorithm to optimize our model.

3) ViewNet Testing: During the testing stage, ViewNet takes a 2D image and 3D camera coordinates as input and predicts the aesthetics score of the image. We have approximately 4 K images to test ViewNet. The predicted aesthetics score will be used in the viewpoint recommendation stage. We select grids and images according to the predicted score, which will be described in Section III-D.

D. Viewpoint Recommendation

Our viewpoint recommendation method consists of the following steps:

1) Split the 3D Space into Grids: We split the 3D space that contains all camera positions of a POI in the x, y and z directions. In the x direction, we find the minimum value of x and maximum value of x among all camera coordinates and denote them as $x_{\min}$ and $x_{\max}$, respectively. Then, we equally split the space between $x_{\min}$ to $x_{\max}$ into 5 segments. The same operation is performed on the y direction and z direction. Thus, the space is split into $5 \times 5 \times 5$ grids, and each grid may or may not contain several images.

2) Calculate the Biased Mean Score of Each Grid: Given the jth image and its camera coordinate (x, y, z) in the ith grid, we use the trained ViewNet to assess the aesthetics s_{ij}. s_{ij} is equal to Z in Equation (4). Then, we calculate the biased mean score of each grid with Equation (6):

$$S_i = \frac{\sum_{j=1}^{N_i} s_{ij}}{N_i + 1} \tag{6}$$

where $s_{ij} \in [1, 4]$ is the aesthetics score of image j in grid i and N_i is the number of images in grid i. The 1 in the denominator is used to avoid the division-by-zero problem.

3) Viewpoint Recommendation: We rank all the grids in descending order according to their biased mean scores S_i. We recommend different grids to maintain diversity since images in different grids are taken from different viewpoints. The top N grids with the highest S_i are recommended in our experiment. Within each grid, we recommend the top K images with

the highest aesthetics, thus ensuring the aesthetics of the recommendation results.

IV. Experiment

To show the effectiveness of our method, we compare our method with other approaches, such as canonical views (CV) [24], clustering, ranking and ranking (CRR) [33], identical semantic points (ISP) [34], high-frequency shooting location (HFSL) [25] and the PSAE [23]. We now give a brief description of these methods:

CV: This is a clustering-based method for scene summarization. It uses image collections from the Internet and examines the distribution of images to select a set of canonical views via visual feature clustering.

CRR: This uses a combination of context- and content-based tools to generate representative sets of images. CRR considers the number of users, but this factor requires a large amount of accurate data.

ISP: This groups landmark images by viewpoint album (VA) generation and expresses the relative viewpoint of an image with a 4D viewpoint vector for the horizontal, vertical, scale and rotation aspects. Then, it summarizes the landmarks in terms of viewpoints.

HFSL: This is an author-topic model-based collaborative filtering method that is used to make recommendations for social media users.

PSAE: The basic idea of this method is to treat viewpoints that are selected by more users as salient areas. Then, it assesses image aesthetics according to the distance between the picture center and the salient area center. This method recommends images based on image aesthetics and diversity.

A. Dataset and Data Preprocessing

1) Filter Out Images Without GPS Information: We crawled 9105 images from Flickr, and these images are from 9 POIs: #1) Arc de Triomphe, #2) Big Ben, #3) Cologne Cathedral, #4) Eiffel Tower, #5) Leaning Tower of Pisa, #6) Mount Rushmore, #7) Statue of Liberty, #8) Taj Mahal, and #9) Tiananmen. Since GPS information is essential in our data preprocessing step, we keep only images with GPS information in our dataset and filter out images without GPS information during the crawling process. Ultimately, we obtained 9105 images of 9 famous POIs.

We invited 25 volunteers to annotate 7562 images. We performed data augmentation on the dataset to obtain a larger dataset of size approximately 20 K. We used 16 K images for training and 4 K images for validation. The data augmentation methods we used include color jitter, random crop, and noise. Each image is assigned an integer score from 1 to 4, and the values [1, 2, 3, 4] represent very bad, bad, good and perfect, respectively. Fig. 7 illustrates some images of different aesthetics.

2) Filter Out Images With Wrong Location: Users may assign a tag to an image randomly. For example, an image tagged with "Statue of Liberty" may have a different location that is not even in New York. Images with wrong locations will cause great difficulty for the feature matching step in 3D reconstruction or even make it fail, so such images must be filtered out.

(a) score = 1 (b) score = 2 (c) score = 3 (d) score = 4

Fig. 7. Example images of different scores. The images from (a) to (b) have aesthetics scores from 1 to 4 respectively.

We calculate the distance of two points $p_1 = (\lambda_1, \phi_1)$ (where λ_1 and ϕ_1 are the longitude and latitude of p_1, respectively) and $p_2 = (\lambda_2, \phi_2)$ with Equation (7).

$$\begin{aligned} d &= r\Delta\sigma \\ \Delta\phi &= |\phi_1 - \phi_2| \\ \Delta\lambda &= |\lambda_1 - \lambda_2| \\ \Delta\sigma &= 2\arcsin\sqrt{\sin^2\left(\frac{\Delta\phi}{2}\right) + \cos\phi_1 \cdot \cos\phi_2 \cdot \sin^2\left(\frac{\Delta\lambda}{2}\right)} \end{aligned} \tag{7}$$

where d is the distance between two points and r is the radius of the Earth. After filtering out images whose geographic distance from the ground truth is greater than 3 km, 7562 images remain.

B. Evaluation Criteria

In this section, we describe the evaluation criteria used. Both the aesthetics and diversity of the recommended results are compared. To make the comparison fair, we invited 25 volunteers to assess the aesthetics and diversity scores of the recommendation results.

Each resulting image for a POI is assigned an aesthetics score $aes_i \in \{1, 2, 3, 4\}$, and these stand for very bad, bad, good and perfect, respectively. We adopt the *mean aesthetics score* of each POI to evaluate the aesthetics of the recommended results:

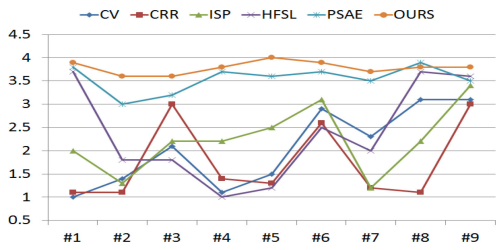
$$mas = \frac{1}{N}\sum_{i=1}^{N} aes_i \tag{8}$$

where N is the number of recommended images of a POI. Every POI image set recommended by a method is assigned a diversity score $div_i \in \{1, 2, 3, 4\}$; these stand for very monotonous, monotonous, diverse and very diverse, respectively.

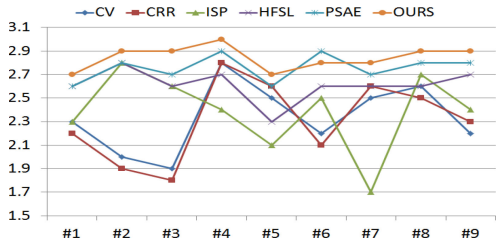
C. Aesthetics Evaluation and Viewpoint Recommendation

We compare our viewpoint recommendation approach with other approaches by objective and subjective performance comparisons.

1) Objective Performance Comparison: We show the aesthetics and diversity score of different methods in Fig. 8. The aesthetics and diversity scores were assigned by 25 invited volunteers.



(a) comparison of average aesthetics score



(b) comparison of diversity score

Fig. 8. Comparison results of CV, CRR, ISP, HFSL, PSAE, and OURS. We have 16 K images for training ViewNet and 4 K images for testing ViewNet.

From Fig. 8(a), we find that our method achieves a better aesthetics score than other approaches on 8 of the 9 POIs. The exception is that on POI#8, the PSAE has a higher aesthetic score than ours, but the difference is subtle. CV only examines the distribution of images and selects a set of canonical views to form the scene summary by clustering methods and does not analyze image aesthetics; its aesthetics and diversity are not good enough. CRR considers the number of users, but this factor requires a large amount of accurate data. Other methods consider the shooting frequency of specific locations, but they do not consider the viewpoint factor.

From Fig. 8(b), we find that the PSAE and our method have the highest diversity score. The reason is that our method recommends viewpoints from different grids; these different grids represent diverse viewpoints, so it can achieve high diversity. The PSAE considers the location of images and the salience information so that top-ranked images are selected from diverse perspectives.

Our model utilizes the powerful feature extraction and expression ability of artificial neural networks, especially CNNs. In addition, with the auxiliary effect of 3D pose information, our model can obtain better validation accuracy, as shown in Table VI.

2) Subjective Performance Comparison: Fig. 9 shows our recommendation results, which are compared with those of CV, CRR, ISP, HFSL and the PSAE. We split the candidate camera pose space into different grids, and each grid's mean aesthetics score is considered. We recommend images from different grids; these grids are standard for different viewpoints, which guarantees the diversity of recommended viewpoints.

The PSAE recommends images of a POI with the basic idea that viewpoints from which more people shoot images are of high aesthetic quality. Fig. 9 shows that other methods may recommend incomplete images. Our method learns 2D image features and 3D pose features jointly, which means that it learns the complete features of a viewpoint. Incomplete images are more likely to be removed by our method, and our method tends to recommend complete images.

D. Discussion

1) Effectiveness of Object Detection in Speeding up 3D Reconstruction: We use our weakly supervised object detection method to crop the main object from the original images.

To show that it is useful to detect the main object of each POI before the SfM step, we perform 3D reconstruction with and without previous weakly supervised object detection. Table III shows the runtime of both cases. In Table III, T_1 is the runtime for 3D reconstruction without weakly supervised object detection, T_{21} is the runtime for weakly supervised object detection, and T_{22} is the runtime for 3D reconstruction after weakly supervised object detection. The total 3D reconstruction time for the 9 POIs is shown in the last row, and T_{21} and T_{22} are added to obtain the total time of weakly supervised object detection and 3D reconstruction. It is obvious that detecting the main object before 3D reconstruction is 2.75 times as fast as the original 3D reconstruction method.

2) Effectiveness of Object Detection in Improving the 3D Reconstruction Quality: In addition to speeding up 3D reconstruction, performing weakly supervised object detection also improves the reconstruction results. Table IV shows the quality comparison of 3D reconstruction without and with weakly supervised object detection. The quality score of 3D reconstruction is assigned by 25 volunteers, and the scores {1, 2, 3, 4} represent very bad, bad, good and very good, respectively. From Table IV, we find that we obtain better 3D reconstruction quality by performing weakly supervised object detection before 3D reconstruction. The reason for the improved results is that weakly supervised object detection removes most irrelevant background keypoints, and these background keypoints are the major source of errors in matches.

Fig. 10 shows some intuitive examples. There are some false points in the bottom-right of the first row in Fig. 10(a) and on the right of the second row in Fig. 10(a). The false points in the first row of Fig. 10(a) even form a spire of the POI *Big Ben*, which is unacceptable because the POI has only one spire.

3) Number of Channels to Select in the Feature Map to Detect Foreground Objects: The number of channels used for detecting foreground objects can affect the detection result to some degree. To compare the performance of using different numbers of feature channels for object detection, we annotate the bounding box of the foreground of some images and apply data augmentation to obtain a dataset of size 10 K. We compare the mean IoUs of methods using different numbers of channels, and the results are shown in Table V.

Table V shows that using 3 channels achieves the best accuracy for detecting foreground objects. Therefore, we set $N = 3$ when we select the top N channels for detecting the foreground main object.

4) Comparison of Clustering-Based Recommendation and Grid-Based Recommendation: In Section III-D, we split the 3D

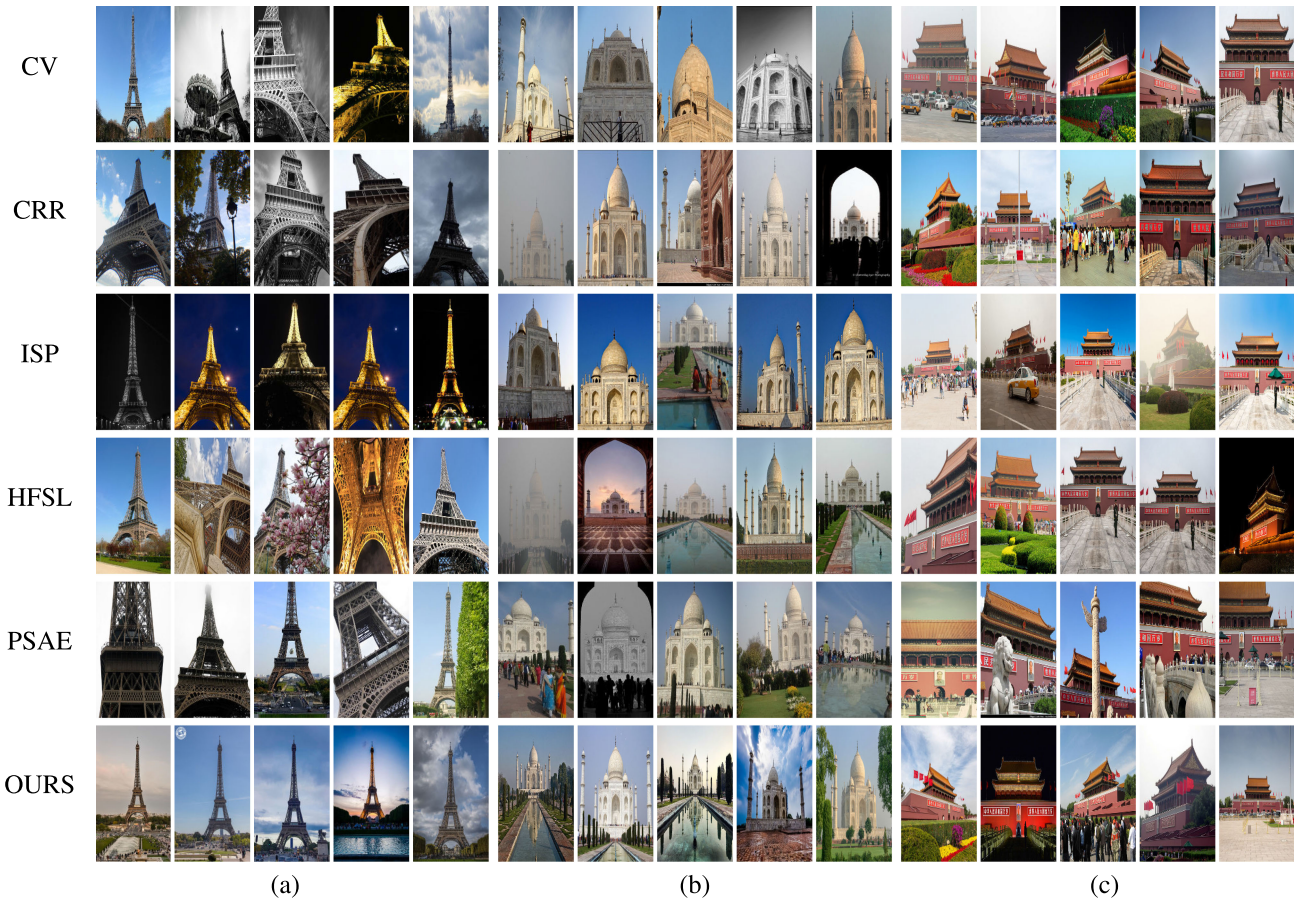


Fig. 9. Recommendation results of CV, CRR, ISP, PSAE, and OURS for three POIs: (a) Eiffel Tower, (b) Taj Mahal, (c) Tiananmen. In our method, the top-5 grids with highest S_i (in Equation (6)) are recommended, within each grid, we recommend top-1 image with highest aesthetics.

TABLE III
RUNTIME COMPARISON OF TWO 3D RECONSTRUCTION METHODS. THE UNIT FOR ALL THE TIME IS MINUTE

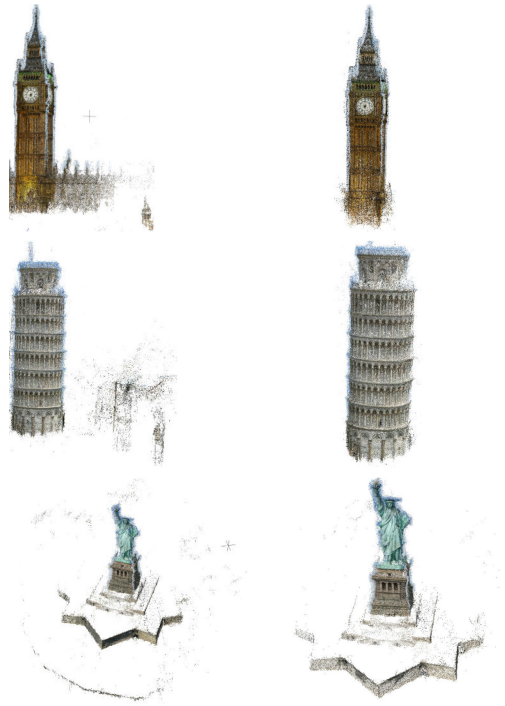
POI	T_1	T_{21}	T_{22}
#1	1672	0.14	788
#2	1550	0.13	570
#3	291	0.07	194
#4	681	0.14	269
#5	360	0.05	145
#6	1433	0.11	449
#7	449	0.12	104
#8	1146	0.13	399
#9	925	0.11	170
total	8507		**3089**

TABLE IV
QUALITY COMPARISON OF 3D RECONSTRUCTION WITHOUT AND WITH WEAKLY SUPERVISED OBJECT DETECTION

POI	$Q_{without}$	Q_{with}
#1 Arc de Triomphe	3	3
#2 Big Ben	1	4
#3 Cologne Cathedral	4	4
#4 Eiffel Tower	3	4
#5 Leaning Tower of Pisa	2	4
#6 Mount Rushmore	3	3
#7 Statue of Liberty	2	4
#8 Taj Mahal	3	4
#9 Tiananmen	3	4
Average	2.7	**3.8**

space that contains all camera positions into grids equally in the x, y and z directions. Another split method is mean-shift clustering. Concretely, we can perform mean-shift clustering on camera coordinates and recommend clusters with high mean aesthetic scores. Fig. 11(b) is a comparison of recommendation aesthetics, and Fig. 11(b) shows a comparison of diversity ('ave' in Fig. 11 is the average score of all the POIs). We find that the aesthetic differences are subtle, while the diversity of grid-based recommendations for some POIs is better. The reason is that the grid-based split method splits the 3D space equally, so recommending different grids can ensure diversity. The clustering-based split method cannot ensure diversity when the camera positions of a POI are almost uniformly distributed, because the clustering algorithm itself cannot work well under such conditions.

5) Effectiveness of Camera Pose Embedding in ViewNet: We now discuss the effect of the 3D camera pose embedding feature on the validation accuracy. We train ViewNet with and without the 3D camera embedding feature and record the validation accuracy on all the POIs. The comparison result is shown in Table VI.

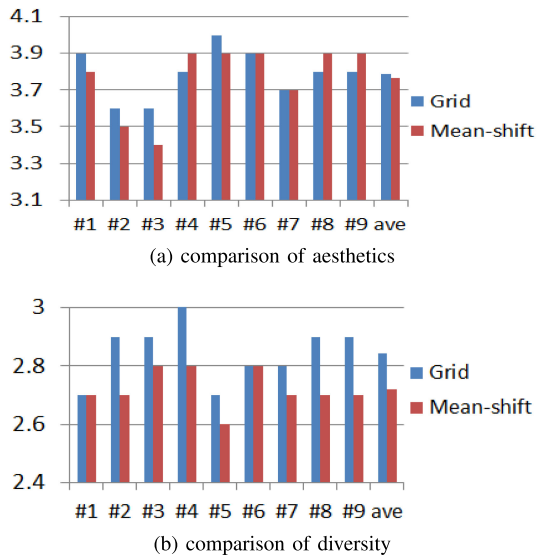


(a) Without object detection (b) With object detection

Fig. 10. Effectiveness of weakly supervised object detection before 3D reconstruction. The POIs from first row to last row are: Big Ben, Leaning Tower of Pisa, Statue of Liberty.

TABLE V
PERFORMANCE COMPARISON OF USING DIFFERENT NUMBER OF CHANNELS. N_c IS NUMBER OF CHANNELS

N_c	1	2	3	4	5	6	7
Mean IoU	0.583	0.615	0.618	0.609	0.598	0.601	0.598



(a) comparison of aesthetics

(b) comparison of diversity

Fig. 11. Comparison of mean shift clustering based recommendation and grid based recommendation.

TABLE VI
VALIDATION ACCURACY COMPARISON OF VIEWNET WITH AND WITHOUT CAMERA POSE EMBEDDING

POI	accuracy without camera pose	accuracy with camera pose
#1 Arc de Triomphe	0.816	0.837
#2 Big Ben	0.814	0.835
#3 Cologne Cathedral	0.770	0.846
#4 Eiffel Tower	0.818	0.818
#5 Leaning Tower of Pisa	0.817	0.833
#6 Mount Rushmore	0.794	0.824
#7 Statue of Liberty	0.821	0.821
#8 Taj Mahal	0.821	0.846
#9 Tiananmen	0.786	0.821
Average	0.812	**0.833**

From Table VI, we find that we obtain better validation accuracy on most POIs and better average validation accuracy with camera pose embedding. We repeat the training and validation 10 times, and the standard deviation of accuracy with and without the camera pose is 0.0146 and 0.0131, respectively. This means that the improvement is not just a chance variation, though the improvement is not large.

V. CONCLUSION

In this paper, we have proposed a novel viewpoint recommendation method based on weakly supervised object detection and 3D reconstruction. By using weakly supervised object detection before 3D reconstruction, we achieve two purposes: 1) the 3D reconstruction speed is 2.75 times as fast as the original speed because we are dealing only with keypoints that lie in the main objects that we are truly interested in. 2) We obtain better reconstruction results, since we remove irrelevant background keypoints. Without the interference of background keypoints, we obtain fewer errors in matching and higher-quality 3D reconstruction results. Our system jointly learns from both 2D image features and 3D camera pose features. The trained model is used for assessing the aesthetic quality of images that are used for 3D reconstruction. Then, the 3D space of all candidate shooting viewpoints is divided into 3D grids, and the average aesthetics of each grid is evaluated by our model. Then, we recommend images from different grids and different grid standards for different viewpoints. By considering both the aesthetic quality and the diversity of possible candidates, our system can recommend several viewpoints that will help users take high-quality photographs.

REFERENCES

[1] X. Qian, X. Lu, J. Han, B. Du, and X. Li, "On combining social media and spatial technology for POI cognition and image localization," *Proc. IEEE*, vol. 105, no. 10, pp. 1937–1952, Oct. 2017.
[2] L. Liu, R. Chen, L. Wolf, and D. Cohen-Or, "Optimizing photo composition," in *Computer Graphics Forum*, vol. 29. Hoboken, NJ, USA: Wiley, 2010, pp. 469–478.
[3] J. Liu, F. Meng, F. Mu, and Y. Zhang, "An improved image retrieval method based on sift algorithm and saliency map," in *Proc. 11th Int. Conf. Fuzzy Syst. Knowl. Discovery*, 2014, pp. 766–770.

[4] K. Lan and X. Qian, "Social image aesthetic measurement based on 3D reconstruction," in *Proc. Int. Conf. Internet Multimedia Comput. Service*, 2014, pp. 350–354.
[5] L. Zhang, Y. Gao, R. Zimmermann, Q. Tian, X. Li, "Fusion of multichannel local and global structural cues for photo aesthetics evaluation," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1419–1429, Mar. 2014.
[6] B. Cheng, B. Ni, S. Yan, and Q. Tian, "Learning to photograph," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 291–300.
[7] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: Exploring photo collections in 3D," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 835–846, 2006.
[8] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 21–37.
[9] J. Li *et al.*, "Attentive contexts for object detection," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 944–954, 2016.
[10] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast R-CNN for pedestrian detection," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 985–996, Apr. 2018.
[11] S. Wang, J. Cheng, H. Liu, F. Wang, and H. Zhou, "Pedestrian detection via body part semantic and contextual information with DNN," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3148–3159, Nov. 2018.
[12] F. S. Saleh, M. S. Aliakbarian, M. Salzmann, L. Petersson, J. M. Alvarez, and S. Gould, "Incorporating network built-in priors in weakly-supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1382–1396, Jun. 2018.
[13] P. Tang *et al.*, "Weakly supervised region proposal network and object detection," *Eur. Conf. Comput. Vision*, 2018, pp. 370–386.
[14] P. Tang *et al.*, "PCL: Proposal cluster learning for weakly supervised object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 176–191, Jan. 2020.
[15] P. Tang, X. Wang, X. Bai, and W. Liu, "Multiple instance detection network with online instance classifier refinement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2017, pp. 3059–3067
[16] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 2846–2854.
[17] O. Chum and A. Zisserman, "An exemplar model for learning object classes," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2007, pp. 1–8.
[18] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[19] Z. Zhou, F. Shi, J. Xiao, and W. Wu, "Non-rigid structure-from-motion on degenerate deformations with low-rank shape deformation model," *IEEE Trans. Multimedia*, vol. 17, no. 2, pp. 171–185, Feb. 2015.
[20] I. Khan, "Robust sparse and dense nonrigid structure from motion," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 841–850, Apr. 2017.
[21] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 4104–4113.
[22] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1362–1376, Aug. 2010.
[23] X. Qian, C. Li, K. Lan, X. Hou, Z. Li, and J. Han, "POI summarization by aesthetics evaluation from crowd source social media," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1178–1189, Mar. 2018.
[24] I. Simon, N. Snavely, and S. M. Seitz, "Scene summarization for online image collections," in *Proc. IEEE 11th Int. Conf. Comput. Vision*, 2007, pp. 1–8.
[25] S. Jiang, X. Qian, J. Shen, Y. Fu, and T. Mei, "Author topic model-based collaborative filtering for personalized poi recommendations," *IEEE Trans. Multimedia*, vol. 17, no. 6, pp. 907–918, Jun. 2015.
[26] X. Qian, M. Li, Y. Ren, and S. Jiang, "Social media based event summarization by user–text–image co-clustering," *Knowl.-Based Syst.*, vol. 164, pp. 107–121, 2019.
[27] X. Qian, Y. Wu, M. Li, Y. Ren, S. Jiang, and Z. Li, "Last: Location-appearance-semantic-temporal clustering based POI summarization," *IEEE Trans. Multimedia*, to be published, doi: 10.1109/TMM.2020.2977478.
[28] J. He, L. Wang, W. Zhou, H. Zhang, X. Cui, and Y. Guo, "Viewpoint assessment and recommendation for photographing architectures," *IEEE Tran. Visualization Computer Graph.*, vol. 25, no. 8, pp. 2636–2649, Aug. 2019.
[29] Y. S. Rawat and M. S. Kankanhalli, "Clicksmart: A context-aware viewpoint recommendation system for mobile photography," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 1, pp. 149–158, Jan. 2017.
[30] S. Lok, S. Feiner, and G. Ngai, "Evaluation of visual balance for automated layout," in *Proc. 9th Int. Conf. Intell. User Interfaces*, 2004, pp. 101–108.
[31] G. Shi, X. Xu, and Y. Dai, "Sift feature point matching based on improved ransac algorithm," in *Proc. 5th Int. Conf. Intell. Human-Mach. Syst. Cybern.*, 2013, vol. 1, pp. 474–477.
[32] R. Arandjelović and A. Zisserman, "Three things everyone should know to improve object retrieval," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2012, pp. 2911–2918.
[33] L. S. Kennedy and M. Naaman, "Generating diverse and representative image search results for landmarks," in *Proc. 17th Int. Conf. World Wide Web*, 2008, pp. 297–306.
[34] X. Qian, Y. Xue, X. Yang, Y. Y. Tang, X. Hou, and T. Mei, "Landmark summarization with diverse viewpoints," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 11, pp. 1857–1869, Nov. 2015.
[35] X. Shi and X. Qian, "Exploring spatial and channel contribution for object based image retrieval," *Knowl. Based Syst.*, vol. 186, 2019, Art. no. 104955.

Ke Li received the B.S. degree from the China University of Mining and Technology, Xuzhou, China, in 2017. He is currently working toward the M.S. degree with the School of Software Engineering, Xian Jiaotong University, Xian, China.

Yuxia Wu received the B.S. degree from Zhengzhou University, Henan, China, in 2014, the M.S. degree from the Fourth Military Medical University, Xi'an, China, in 2017. He is currently working toward the Ph.D. degree with Xi'an Jiaotong University, Xi'an. Her research interests include osocial multimedia mining and recommender systems.

Yao Xue received the B.S. degree from the Xi'an University of Posts & Telecommunications, Xi'an, China, in 2010, the M.S. degree from Xi'an Jiaotong University, Xi'an, in 2013, and the Ph.D. degree from the University of Alberta, Edmonton, AB, Canada, in 2018. He is currently a Researcher with Xi'an Jiaotong University. His research interests include computer vision, medical image analysis, machine learning, and artificial intelligence.

Xueming Qian (Member, IEEE) received the B.S. and M.S. degrees from the Xi'an University of Technology, Xi'an, China, in 1999 and 2004, respectively, and the Ph.D. degree from the School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, in 2008. He was a Visiting Scholar with Microsoft Research Asia from 2010 to 2011. He was an Assistant Professor with Xi'an Jiaotong University, where he was an Associate Professor from 2011 to 2014, where he is currently a Full Professor. He is also the Director of the SMILES Laboratory, Xi'an Jiaotong University. His research interests include by the National Natural Science Foundation of China, Microsoft Research, and the Ministry of Science and Technology. His research interests include social media big data mining and search. He received the Microsoft Fellowship in 2006 and the Outstanding Doctoral Dissertations of Xi'an Jiaotong University and Shaanxi Province, in 2010 and 2011, respectively.