A-Lamp: Adaptive Layout-Aware Multi-Patch Deep Convolutional Neural

Network for Photo Aesthetic Assessment

|  |  |
| --- | --- |
| First Author  Institution1  Institution1 address firstauthor@i1.org | Second Author  Institution2  First line of institution2 address secondauthor@i2.org |

Abstract

*Deep convolutional neural networks (CNN) have recently been shown to generate promising results for aesthetics assessment. However, the performance of these deep CNN methods is often compromised by the constraint that the neural network only takes the fixed-size input. To accommodate this requirement, input images need to be transformed via cropping, wrapping, or padding, which often alter image composition, reduce image resolution, or cause image distortion. Thus the aesthetics of the original images is impaired because of potential loss of fine grained details and holistic image layout. However, such fine grained details and holistic image layout is critical for evaluating an images aesthetics. In this paper, we present an Adaptive Layout-Aware Multi-Patch Convolutional Neural Network (A-Lamp CNN) architecture for photo aesthetic assessment. This novel scheme is able to accept arbitrary sized images, and learn from both fined grained details and holistic image layout simultaneously. To support A-Lamp training on these hybrid inputs, we extend the method by developing a dedicated double-subnet neural network structure, i.e. a MultiPatch subnet and a Layout-Aware subnet. We further construct an aggregation layer to effectively combine the hybrid features from these two subnets. Extensive experiments on the large-scale aesthetics assessment benchmark (AVA) demonstrate significant performance improvement over the state of the art in photo aesthetic assessment.*

# Introduction

Automatic image aesthetics assessment has drawn numerous research attentions with the goal of endowing computers with the capability of perceiving aesthetics and visual quality as humans. Potential usage for this emerging technology could be found in a wide range of contemporary applications from advanced artificial intelligent systems to real-time, consumer mobile applications.

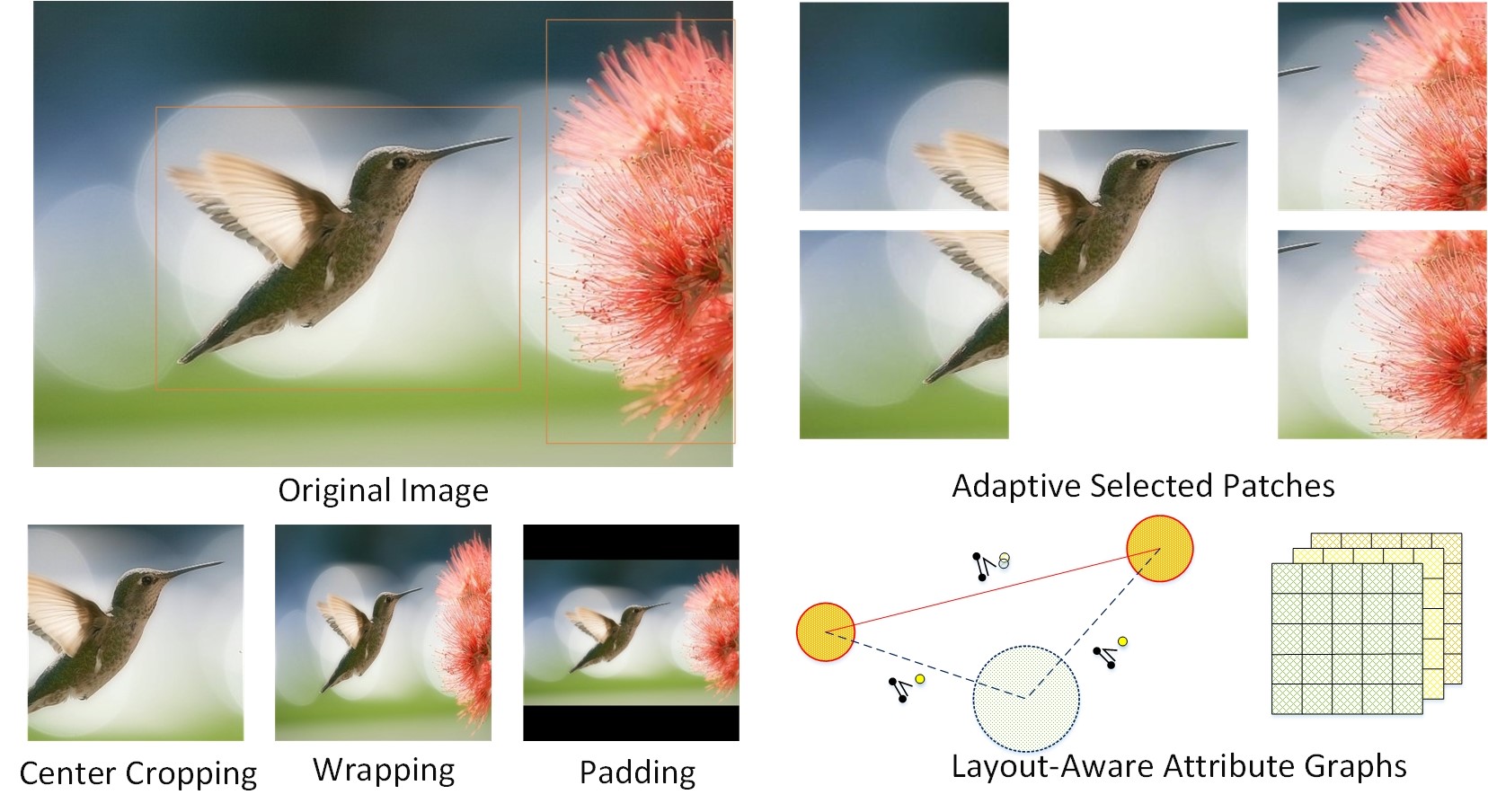


Figure 1. Illustration fo image transformation.

However, assessing photo aesthetics by machines is challenging. Over a decade has passed since the initial attempt. Among the early efforts [6, 16], various hand-craft aesthetics features have been manually designed to approximate a number of photographic and psychological aesthetics rules. These include low level features [26, 2] based on the distribution of edges, color histograms and light contrast, and so on, as well as high level features [39, 7, 37, 5] based on composition principles, e.g. ”Rule of Thirds”, ”Visual Balance” and ”Golden Ratio”, Low-of Depth, color harmony, and photo content and scene categories. Although these handcraft features have shown encouraging results, to design effective aesthetics features manually is still a challenging task because even the very experienced photographers use very abstract terms to describe high quality photos. Other approaches have also been developed to leverage more generic image features, such as SIFT, Fisher Vector [28, 33] and bag of visual words [37], to predict photo aesthetics. Although very good performance has been achieved, the image representation provided by these generic features may not be suitable for the task of photo aesthetics, as they are designed to capture the general semantics of the natural images instead of describing the aesthetics of the image composition.

Because of the limitations of image feature-based approaches, many researchers have recently turned to deep learning strategy to extract effective aesthetics features automatically from training data [24, 22, 27, 38, 15]. These deep CNN methods have indeed shown promising results. However, the performance is often compromised by the constraint that the neural network only takes the fixed-size input. To accommodate this requirement, input images will need to be obtained via cropping, wrapping, or padding. These additional operations on the original images often alter image composition, reduce image resolution, or cause extra image distortion, and thus impair the aesthetics of the original images because of potential loss of fine grained details and holistic image layout. It has been long recognized in aesthetics assessment community that such fine-grained details and overall image layout have been intrinsically utilized by professional photographer to carry out photo quality assessment and have been incorporated in modern image related applications in image quality estimation [14, 3], image aesthetics categorization [22, 30, 20, 32, 40, 46], and image style classification [15, 9]. As we can see from Fig.1, one randomly cropped patch may generate ambiguity in training examples because the aesthetic attributes in one patch may not adequately represent the holistic information in the entire image (the flower is eliminated and only the bird is retained). Uniformly wrapping also reduces original image resolution and distorts the salient object and thus compromises the detailed clarity of the important regions. The artificial boundaries between the original image and the padding area may also confuse the neural network. Training from such artificially transformed images does not capture the essence of the aesthetics in each of the input image and thus compromise the ability of the network to effectively learn the desired discriminative features.

The limitation in fixed-size input has been addressed by training images in a few different scales to mimic varied input sizes [10, 27]. However, they still learn from transformed images, which may still result in substantial loss of fine grained details and undesired distortion of image layout. To enable learning from fine grained details, a deep multi-patch aggregation network architecture (DMA-Net) was proposed in [24] to take multiple random cropped patches as the input. This scheme shows some promising results. However, these randomly picked bag of patches lacks of efficiency and cannot represent the overall image layout. In general, this random cropping strategy requires a large number of training epochs to cover the desired diversity in training, which lead to very low efficiency in learning.

To resolve the these two important issues, we present an

Adaptive Layout-Aware Multi-Patch Convolutional Neural Network (A-Lamp CNN) architecture to accomplish learning-based photo aesthetic assessment. This dedicated CNN can accept arbitrary image sizes, and deal simultaneously with issues resulting from both the fine-grained details as well as image layout. Learning from fine grained details is achieved by constructing multiple, shared columns in a Multi-Patch subnet and feeding multiple patches to each of the columns. Instead of random cropping adopted in the existing scheme, we propose an adaptive multi-patch selection scheme to enhance the training efficiency and achieve significant performance improvement. In addition, we also designed a separate layout aware subnet in order to learn from the holistic image composition by representing local and global layout of the images through leveraging attribute graphs. An aggregation layer is then adopted to effectively combine the hybrid features from two subnets. We shall carry out extensive experiments on the large-scale aesthetics assessment benchmark (AVA) to demonstrate the effectiveness of the proposed architecture for photo aesthetic assessment.

1.1. Challenges and Contribution

Learning both fine-grained details and image layout is indeed very challenging. First, the detail information of an image is contained in the original high resolution images. Training deep networks with large-sized input dimensions requires much longer training time and significantly larger network structure, training dataset, and hardware memory. To design a practical learning scheme to capture the fine-grained details, we formulate the problem by representing an input image with a set of carefully cropped patches and associating this set with the image’s annotation labels. An aggregation structure that leverages statistical functions is adopted to incorporate these multiple patch instances. In order to enhance the training efficiency, we propose an adaptive multi-patch selection strategy which is fundamentally from the existing random cropping method reported in [24]. The central idea of multi-path selection is to maximize the input information more efficiently. We achieve this goal by dedicatedly selecting the patches that play important role in affecting images’ aesthetics. We expect that the proposed strategy shall be able to outperform the random cropping scheme even with substantially less training epochs.

Second, how to effectively describe specific image layout and incorporating it into the deep CNN is again very challenging. Existing works related to image layout descriptors are dominantly based on a few simple photography composition principles, such as visual balance, rule of thirds, golden ratio, and so on. However, these general photography principles are inadequate to represent local and global image layout variations. To incorporate global layout information into CNN, transformed image inputs via wrapping and center-cropping have been used to represent the global view [23]. However, such global transformation often alters the original image composition or may cause undesired layout distortion.

In this research, we represent various input images’ layout by constructing attribute graphs. We use graph nodes to represent objects and the global scene in the image. Each object (note) is described using object-specific local attributes while the overall scene is represented with global attributes. The combination of both local and global attributes captures the layout of an image effectively. This attribute-graphs based approach is expected to model image layout more accurately and outperform the existing approaches based on warping and center-cropping.

The main contributions of this proposed A-Lamp scheme can be summarized into three-fold:

* We introduce a new neural network architecture to support learning from any image sizes without being limited to small and fixed size of the image input. This shall open a new avenue of deep learning research on arbitrary image size for training.
* We design two novel subnets to support learning at different levels of information extraction: fine-grained image details and holistic image layout. Aggregation strategy is developed to effectively combine hybrid information extracted from individual subnet learning.
* We have also developed an adaptive patch selection strategy to enhance the training efficiency associated with variable size images being used as the input. This aesthetics driven selection strategy can be extended to other image analysis tasks with clearly-defined objectives.

# Related Work

## Deep Convolutional Neural Networks

Recently, deep learning methods have shown great successes in various computer vision tasks, including conventional tasks in object recognition [42], object detection [10, 21], and image classification [34, 11] as well as contemporary tasks in image captioning [1], saliency detection [31], style recognition [9, 15] and photo aesthetics assessment [22, 24, 38, 27, 14]. Most existing deep learning methods transform input images via cropping, scaling, and padding to accommodate the deep neural network architecture requirement in fixed size input which would compromise the network performance as we have discussed previously.

Recently, new strategies to construct adaptive spatial pooling layers have been proposed to alleviate the fixed-size restriction [10] [27]. In theory, these network structures can be trained with standard back-propagation, regardless of the input image size. In practice, the GPU implementations of deep learning are preferably run on fixed input size. The recent research [10] [27] mimic the variable input sizes by using multiple fixed-size inputs which are obtained by scaling from original images. It is apparently still far from arbitrary size input. Moreover, the learning is still from transformed images, which inherently compromise the performance of the deep learning networks.

Others have proposed dedicated network architectures. A double-column deep convolutional neural network was developed in [22] to support heterogeneous inputs with both global and local views. The global view is represented by padded or wrapped image while the local view is represented by randomly cropped single patch. This work was further improved in [24], where a deep multi-patch aggregation network was developed (DMA-Net) to take multiple randomly cropped patches as input. This network have shown some promising results. However, these random order of bag of patches is unable to capture image layout information, which is crucial in image aesthetics assessment. Furthermore, to ensure that most of the information will be captured by the network, this scheme uses large number of randomly selected groups of patches for each image, and trains them for 50 epochs, resulting in very low training efficiency.

## Image Layout Representation

To represent holistic image layout, existing works [20, 30, 32, 40, 44] adopt dominantly the model of image composition by approximating some simple traditional photography composition guidelines, such as visual balance, rule of thirds, golden ratio, and diagonal dominance. However, these heuristic guidance-based descriptors cannot capture the intrinsic of photo aesthetics in terms of image layout.

Attribute-graph, which has long been used by the vision community to represent structured groups of objects [8, 25, 13, 35], shows promising results in representing complicated image layout. The spatial relationship between a pair of objects was considered in [19] even though the overall geometrical layout of all the objects and the object characteristics cannot be accounted for with this method. The scheme reported in [41] was able to maintain spatial relationships among objects but related background information and object attributes were not addressed. The scheme reported in [18] considers both objects and their interrelations, but have not been integrated with the holistic background modeling. The scheme in [4] performs image aesthetics ranking by constructing the triangular object structures with attribute features. However, this scheme lacks of proper account for the global scene context.

To resolve the technical issues associated with existing approaches in image aesthetics assessment, we present in this paper a dedicated CNN architecture named A-Lamp. This novel scheme of A-Lamp can accept arbitrary images with its native size. Training and testing can be effectively performed by considering both fine-grained details and image layout, thus preserving the information from the original images. The design of the proposed A-Lamp CNN is inspired jointly by the success of fine-grained detail learning using multi-patch strategy [24, 21] and the success of holistic layout representation by attribute graph. However, the proposed scheme can successfully overcome the stringent limitations of the existing schemes. Like DMA-Net in [24], this new scheme also crops multiple patches from original images to preserve fine-grained details. Comparing to DMA-Net, this scheme has two major innovations. First, instead of cropping patches randomly, we propose an adaptive patch selection strategy that selects patches that are semantically important. Second, unlike the DMA-Net that just focus on the fine-grained details, this A-Lamp CNN incorporates the holistic layout via the construction of attribute graph. These two innovations result improvement in both efficiency and accuracy over DMA-Net.

# Adaptive Layout-Aware Multi-Patch CNN

The architecture of our proposed A-Lamp is showed in Fig. 3. Given an arbitrary sized image, multiple patches are carefully selected by the *adaptive patch selection* module, and are fed into the *Multi-Patch subnet*. A statistic aggregation layer is followed to effectively combine the extracted features from these multiple patches. At the same time, a trained CNN is adopted to detect objects in the image. The local and global layout of the input image are further represented by an attribute-graph. At the end, a learned aggrega-

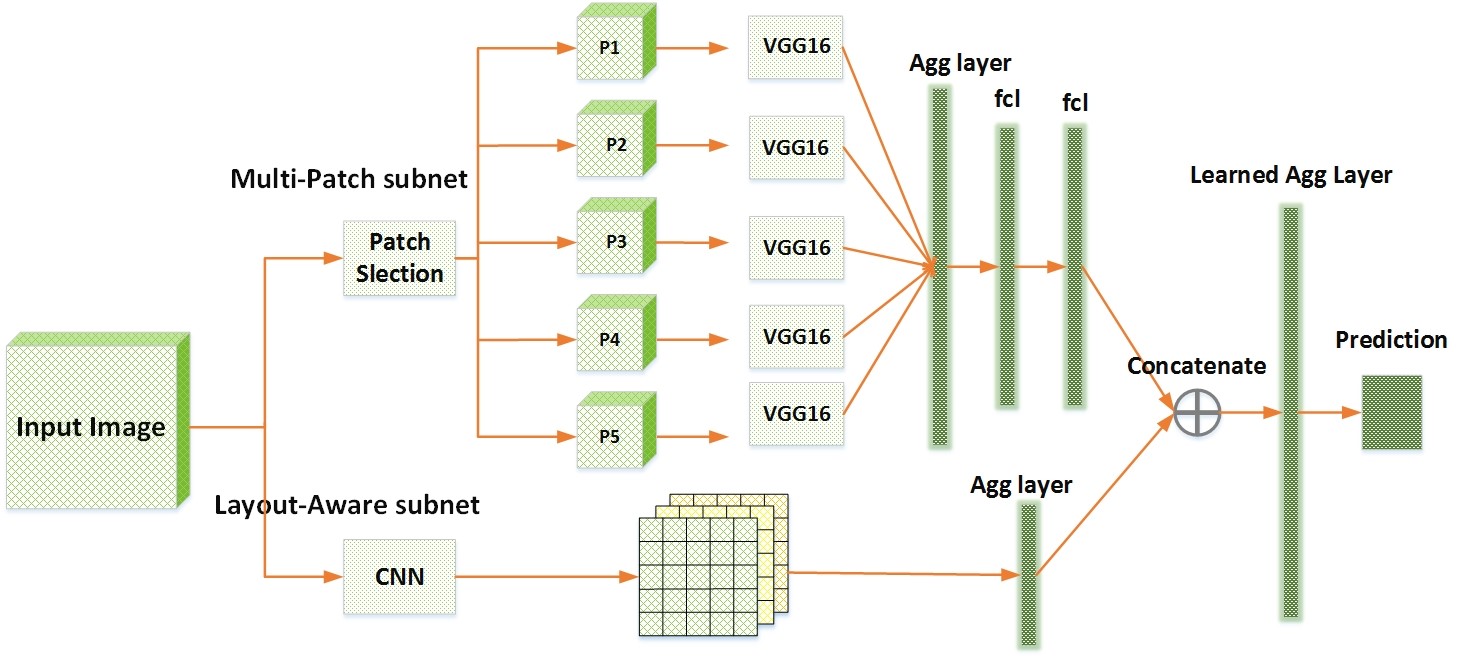


Figure 2. The architecture of the A-Lamp

tion layer is utilized to incorporate the hybrid deep features from the two subnets and finally give the aesthetic prediction. More details will be illustrated in this section.

## Multi-Patch subnet

To support training from fine-grained details, we represent each image with a set of carefully cropped patches, and associate the set with the image’s label. The training data is {P*n,*y*n*}*n*∈[1*,N*], where *Pn* = {*pnm*}*m*∈[1*,M*] is the set of *M* patches cropped from each image. As shown in Fig.3.1, the Multi-Patch subnet contains mainly three parts: an adaptive patch selection module, a set of paralleled CNNs that are used for extracting deep features from each of the patch, and an orderless aggregation structure which combines the extracted deep features from the multicolumn CNNs jointly.

### Adaptive Patch Selection

Instead of randomly cropping 50 sets of patches (totally 250 patches for each image) [24], we aim to carefully select the most discriminative and informative patches to enhance training efficiency. To realize that, we studied professional photography rules and human visual principles. We find that human visual attention does not distribute evenly within an image. That means, some regions play more important roles when people valuating photos, while the others do not.

In addition, holistic analysis is critical for evaluating an image’s aesthetics. Thus just focus on the subjects is not enough for aesthetic assessment. Therefore, several issues are concerned when we perform patch selection:

Saliency Map The task of saliency detection is to identify the most important and informative part of a scene. Saliency map models human visual distribution, and is capable of highlighting visually attention region. Therefore, it is natural to adopt saliency map for selecting regions that human pay much more attention to.

Pattern Diversity Despite that, we also encourage diversification within a set of patches. Different from some conventional computer vision tasks (e.g. image classification and object recognition) which just need to focus on the foreground objects. To evaluate image aesthetics, holistic analysis is critical. Because some important aesthetic characters, e.g. Low-of-Depth, color harmonization, simplicity, etc., are perceived by analyzing both the foregroud and the background as a whole.

Overlapping Constraint Spatial distance for each patch pair is considered to constrain the overlapped ratio of these selected patches.

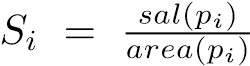
Therefore, we formulate the patch selection as an optimization problem. An objective function is defined to search for the optimal combination of patches:

 *M M M* 

{*c*∗} = *argmax*X*Si* + X*Dp*(Ne*i,*Ne*j*) + X*Ds*(*ci,cj*)

*i,j*∈[1*,M*] *i*=1 *i*6=*j i*6=*j*

(1)

where  is the optimal set of *M* selected patches centers. where {*c*∗} is the optimal subset of selected patch centers.  is the normalized saliency value that each patch *pi* occupies. The saliency value is obtained by a graph-based saliency detection approach [43]. *Dp*(·) is the pattern distance function which measures the difference of two patches’ patterns. Here we adopt edge and chrominance distribution to represent the pattern of each patch. To measure the difference of edge distribution of each patch pairs, we model the edge and chrominance distribution of a patch using a multi-variant

Gaussian. *N*˜*m* = {*Ne*(*µe,*P*e*)*,Nc*(*µc,*P*c*)}*m*∈[1*,M*] denotes the edge distribution *Ne*(*µe,*P*e*) and chrominance distribution *Nc*(*µc,*P*c*) of patch *pm*. We formulate the problem by mapping these distribution to the *Wasserstein*

*Metric space MM*×*M*. The we define the *pth Wasserstein distance* between the distribution *N*˜*i* and *N*˜*j* as:

!1*/p*

Z

*Dp*(*N*˜*i,N*˜*j*) = inf *d*(*x,y*)*pdγ*(*x,y*)

*r*∈Γ(*N*˜*i,N*˜*j*) *M*×*M*

(2) *Ds*(·) is the spatial distance function, which is measured by Euclidean Distance.

### Orderless Aggregation Structure

We perform aggregation of the multiple instances to support our network learn from multiple patches cropped from the original image. Let hBlob be the set of patch features extracted from *nth* image at *lth* layer of the shared CNNs. Where *bni* is a *K* dimensional vector. *Tk* denotes the set of values of the *kth* component of all *bi* ∈

*Blob*, i.e. *Tk* = {*dik*}*i*∈[1*,M*].

|  |
| --- |
| Figure 3. The Multi-Patch subnet |

The orderless aggregation layer is comprised of a collection of statistical functions, i.e., *FAgg* = {*FAggu* }*u*∈[1*,U*]. Each *FAggu* computes *Blob* returned by the shared CNNs. In our work, we propose to have *U* = {*max, mean*}. The outputs of the functions in *U* are concatenated to produce a *Kstat*-dimensional feature vectors. Two fully connected layers are followed for implementing of multi-patch aggregation component. The whole structure can be expressed as a function *f* : {*Blob*} → *Kstat*:

 (3)

where ⊕ is a vector concatenation operator which produces a column vector, *W* ∈ *Kstat*×*UK* is the parameters of the fully-connected layer. Fig. 3.1 shows an example of Statistics Aggregation Structure with *M* = 5 and *K* = 3. In practice, the feature dimension *K* = 4096.

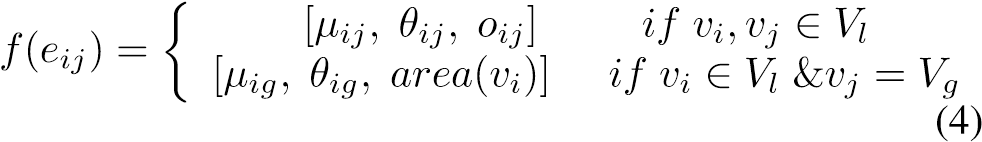
## Layout-Aware Subnet

The layout of an image is another critical ingredient that affecting aesthetics. Good and balanced composition could make an image look more appealing even if the scene being shot is normal. In this paper, we develop a novel Layout-aware subnet, and combine it with the Multi-Patch subnet to effectively enhance the performance of our proposed A-Lamp CNN. To effectively represent image layout, we construct *Attribute-graph*, which is an undirected fully connected graph, incorporating both local and global image characteristics. As can be seen from Fig. 3.2, the graph nodes characterize salient objects as well as the overall scene context using node attributes, while the edges capture the object topology.

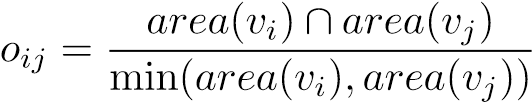
We first employ a trained CNN [45] to localize the salient objects. Let I : {B*i,si*}*Nobj* denotes a set of detected objects in image I, where each object is labeled by a bounding box B*i* and associated with a confidence score *si*, *Nobj* denotes the number of objects. Here *G*(*V,E*) is an undirected fully connected graph. *V* = {*Vlocal,Vglobal*} represents the nodes and *E* represents the set of edges connecting the nodes. Each object present in the image contributes to a graph node resulting in a total of *Nobj* local nodes

*Vlocal* = {*v*1*,*· · ·*,vNobj*}. The global node *Vglobal* represents the background. An image with *Nobj* objects is thus transformed into a graph having *Nobj* +1 nodes. We define two kind of edges, i.e. local edges and global edges. Where local edges refer to the edges between two local nodes, there will be (*Nobj* −1)! such edges. The edges connecting local nodes and global node are global edges, there will be *Nobj* such edges.

Each local node is represented using local attributes. These local attributes are limited to the area occupied by the bounding box of that particular object. The global node captures the overall essence of the image. The features of the model are defined so as to capture the spatial configuration of the image components. The local features capture the relative arrangement of the objects with respect to each other while the global features define the positioning of the objects in the image. The features are represented by



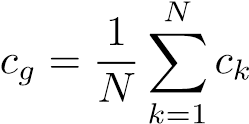
*eij* represents the edge connecting node *vi* to node *vj*. *µij* is the spatial distance between object centroids. *θij* represents the angle of the graph edge with respect to the horizontal taken in the anti-clockwise direction. It indicates the relative spatial organization of the two objects. *oij* represents the amount of overlap between the bounding boxes of the two objects and is given by

 (5)

where *area*(*vi*) is the fraction of the image area occupied by the *ith* bounding box. The intersection of the two bounding boxes is normalized by the smaller of the bounding boxes to ensure the overlap score of one, when a smaller object is inside a larger one. *µig* and *θig* are the magnitude and orientation of the edge connecting the centroid of the object corresponding to node *vi* to the global centroid. The global

|  |
| --- |
| Figure 4. Architecture of the Layout-Aware subnet |

centroid is computed by:

 (6)

where *ck* represents the centroid of the *kth* local node. The global centroid represents the center of the geometrical layout of the objects in the image. The edges connecting each object to the global node illustrate the placement of that object with respect to the overall object topology.

The architecture of our proposed *Layout-Aware subnet* is shown in Fig. 3.2. As we can see, an aggregation layer is adopted to concatenate the constructed attribute graphs into a feature vector *~ν*, and further combined with the *MultiPatch subnet*, which can be seen in Fig. 3. By statistical study, we find that, the confidence score is very low when *Nobj* ≥ 5. Therefore, we set *Nobj* = 4 to fix the dimension of the feature vector *~ν*.

# Implementation Details

In our implementation, we simplify the computational complexity and release the memory burden in training by first training the Multi-Patch subnet and then combining the layout-aware subnet to fine-tune A-Lamp. The weights of multiple shared column CNNs in the Multi-Patch subnet are initialized by the weights of VGG16 which is pre-trained on the ImageNet [17]. VGG16 is one of the state-of-theart object-recognition networks that has been adopted with great success to many different computer vision problems. Following [24], The number of patches in a bag is set to be 5. The patch size is fixed to be 224 ×224 × 3. The base learning rate is 0.01, the weight decay is 1e-5 and momentum is 0.9. All the network training and testing are done by using the Caffe deep learning framework[12]. The networks are trained with the Adam.

# Experiments

We systematically evaluate our method on the AVA dataset [29], which, to our best knowledge, is the largest publicly available aesthetic assessment dataset. The AVA dataset provides about 250,000 images in total. The aesthetics quality of each image in the dataset was rated on average by roughly 200 people with the ratings ranging from one to ten, with ten indicating the highest aesthetics quality. For a fair comparison, we use the same partition of training data and testing data as the previous work [22, 24, 27, 29] ( roughly 20,0000 images for training and 19,000 images for testing). We also follow the same procedure as the previous work to assign a binary aesthetics label to each image in the benchmark. Specifically, images with mean ratings smaller than equal to 5 are labeled as low quality and those with mean ratings larger than 5 are labeled as high quality.

## Comparison with the state-of-the-art

We denote our *Multi-Patch subnet* as Ours-MP-Net and *Layout-Aware Multi-Patch CNN* as A-Lamp. To evaluate the proposed approach, both Ours-MP-Net and A-Lamp are compared with several state-of-the-arts in this section.

### Analysis of adaptive Multi-Patch subnet

For a fair comparison, we first perform training and testing only using our proposed Multi-Patch subnet, and evaluate Ours-MP-Net with the Deep Multi-Patch-Aggregation Network(DMA-Net) in [24]. DMA-Net is a very recent dedicated deep Multi-Patch CNN for aesthetic assessment. Specifically, DMA-Net performs multi-column CNN training and testing. Five randomly cropped patches from each image was used as training, and the label of the image is associated with the bag of patches. DMA-Netave and DMA-Netmax train deep multi-patch aggregation network using standard patch pooling scheme, where DMA-Netave performs average pooling and DMA-Netmax performs max pooling. The DMA-Net using Statistics Aggregation Structure as DMA-Netstat and Fully-Connected Sorting Aggregation Structure as DMA-Netfc.

The results are shown in Table 5.1.1. We can see that, our methods outperform all kinds of DMA-Net ar-

|  |  |
| --- | --- |
| DMA-Netave | 73.1 % |
| DMA-Netmax | 73.9 % |
| DMA-Netstat | 75.4% |
| DMA-Netfc | 75.4% |
| Ours-MP-Net | 81.7% |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Method | Accuracy |  | Method | Accuracy | F-measure |

Table 1. Performance of Multi-Patch subnet

chitectures. Although [24] randomly copped 50 groups of patches, which takes totally 250 patches for each image, and trained the DMA-Net for 50 epochs. The random cropping strategy may lost much useful information and made the training data confusing for network. Instead of random cropping, we adaptively select the most informative and discriminative patches as input, which shows great improvement. From Fig.1, we can see that, the salient objects, i.e. the bird and the flower, are selected. Within these patches, the most important information and the fine-grained details are remained. Despite that, the background, i.e. the blue sky and the green ground, are also selected so that global characteristics, e.g. color harmony, Low-of-Depth, can be learned. More examples of selected patches are shown in Fig. 5.1.1. We can see that, our strategy is not only effective in selecting the most salient regions (e.g. the human’s eyes and face, the orange flowers, etc.), but also capable of encouraging the pattern diversity (e.g. the green leaf and green beans, the flower and the gray wall). What’s more, our adaptive patch selection strategy much enhanced the training efficiency. The result of Ours-MP-Net is obtained by taking 20-30 training epochs, which is much less than 50 epochs in [24], while showing better performance. The reason of higher accuracy may also lies in different CNN architecture we used. [24] adopt shallow CNN, which has only 4 convolution layers and followed by two fully connected layers. We use VGG16 [36], which shows significant improvement.

### A-Lamp CNN Performance

Table 5.1.2 reports results of our A-Lamp CNN on the AVA dataset for image aesthetics categorization. AVA [29] provides the state-of-the-art result for methods that use manually designed features and generic image features for aesthetics assessment. It is obvious that, all the deep CNN methods outperformed the conventional approach.

To examine the effectiveness of our proposed methods, we compare Ours-MP-Net and A-Lamp with the baseline methods which take fixed-size inputs. In particular, we experiment with three VGG16-based aesthetics assessment methods, each operating on a different type of transformed input.

VGG16-Crop: The input of the network is obtained by

|  |  |  |
| --- | --- | --- |
| AVA | 67.0 % | na |
| VGG-Crop | 71.2 % | 0.83 |
| VGG-Scale | 73.8 % | 0.83 |
| VGG-Pad | 72.9 % | 0.83 |
|  | 76.0 | 0.84 |
|  | 77.1 | 0.85 |
|  | 77.4 | *NA*∗ |
|  | 71.2 | *NA*∗ |
|  | 73.25 | *NA*∗ |
|  | 75.4 | *NA*∗ |
| Ours-MP-Net | 81.7% | 0.91 |
| A-Lamp | 82.5 % | 0.92 |

Table 2. A-Lamp CNN performance comparisons with the state-

of-the-art

randomly cropping from the original image with a 224×224 cropping window. This cropping window size is the fixed size required by the VGG16 architecture. During training, we extract five random crops for each image in the training set and train the network on all the crops with their corresponding aesthetics labels. For each testing image, we follow the previous work [24] to predict the aesthetics quality for 50 random crops obtained from the image and take their average as the final prediction result.

VGG16-Scale: The input of the network is obtained by scaling the original input image to the fixed size of 224×224. Both training and testing are conducted on the scaled version of the input images.

VGG16-Pad: The original image is uniformly resized such that the larger dimension becomes 224 and the aspect ratio is preserved. The 224×224 input is then formed by padding the remaining dimension of the transformed image with zero pixels.

We can see that, both our proposed Multi-Patch subnet and the A-Lamp net outperforms these fixed-size input VGG nets. Such results confirmed that training network on multiple patches generates better prediction performance than networks training on a single patch.

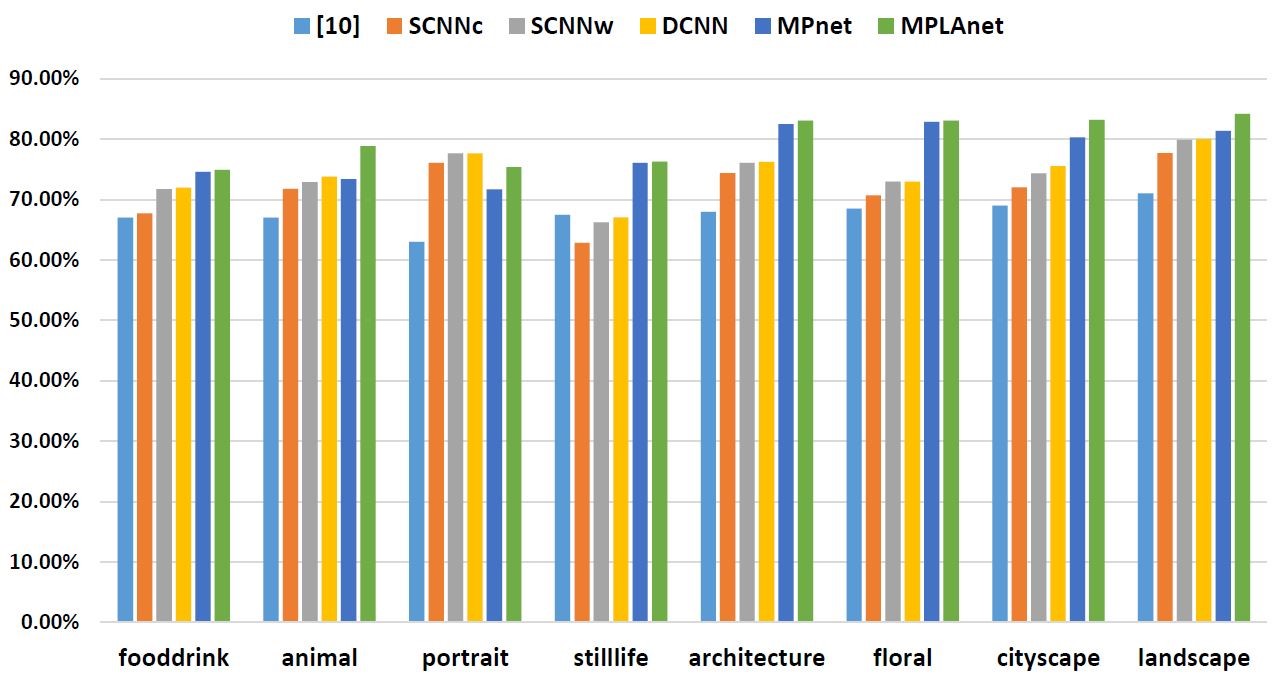
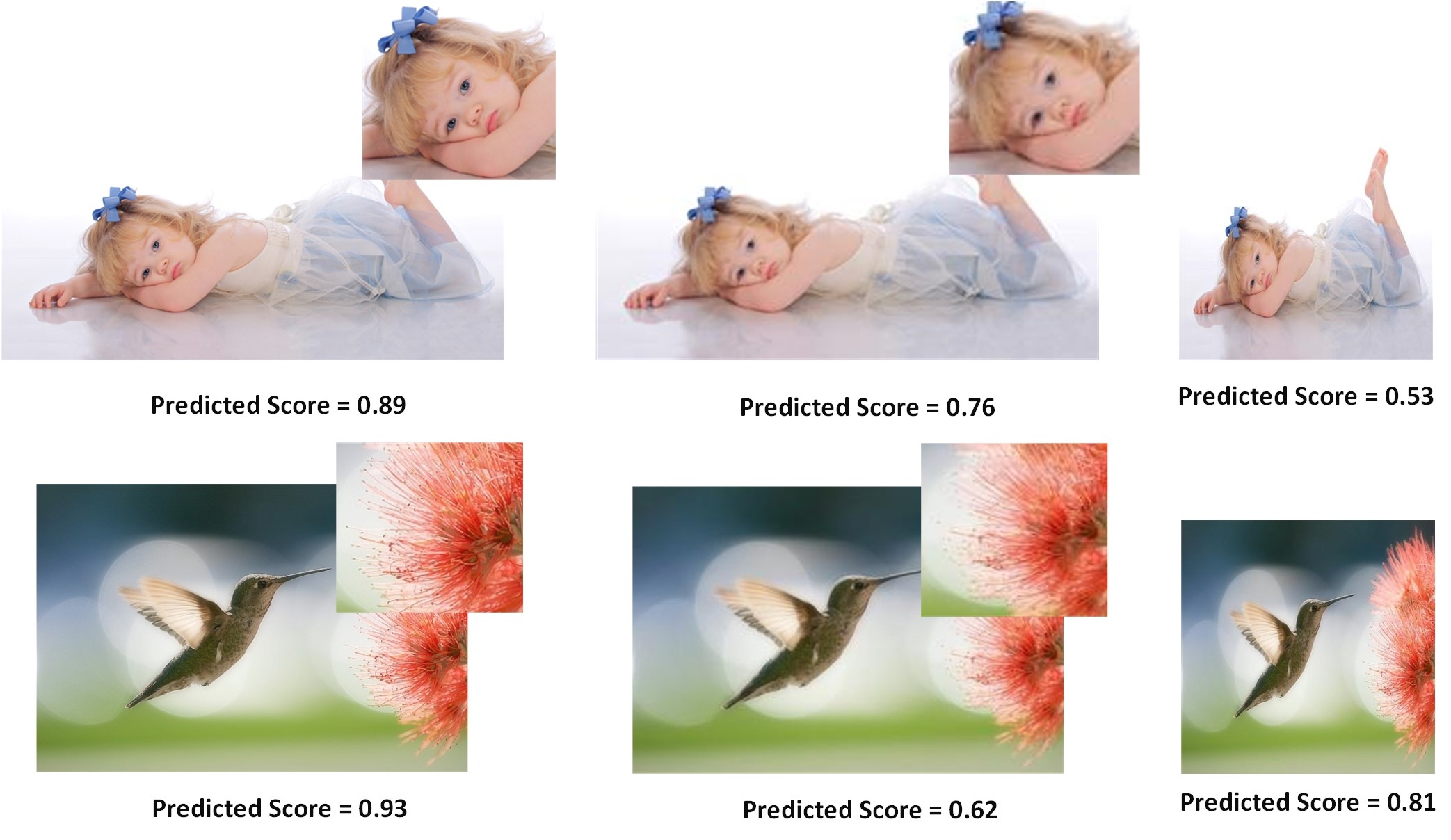
We also compared our work with some latest non-fixedsize restriction methods, i.e. SPP-CNN [10] and MNACNN [27]. Different from these methods that train from several different level of scaled images, we implement the A-Lamp network to be trained from original images. The result turns out learning from original images is critical for aesthetic assessment, as discussed in 1. In addition, higher prediction accuracy further proves that, our proposed adaptive Multi-Patch strategy is more efficient than the spatial pooling layers adopted in SPP-CNN and MNA-CNN.

To show the effectiveness of our proposed layout-aware subnet, we compare A-Lamp with several latest deep CNN networks that incorporate global information for learning.

MNA-CNN-Scene [27] replace the average operator in the MNA-CNN network with a new aggregation layer that takes the concatenation of the sub-network predictions and the image scene categorization posteriors as input and output the final aesthetics prediction. We can see that, the performance not show much improvement after incorporating with scene attributes.

DCNN is a double column convolutional neural network, which allows network training using two inputs extracted from different spatial scales of one image. Specifically, [23] combine random cropped and wrapped images as inputs to train the proposed double-column network. By comparing our A-Lamp test accuracy (82.5 %) with DCNN (73.25 %), we observe that using random cropped and wrapped image to capture local and global image characters is not as effective as our approach.

DMA-Net-ImgFu remains the global view of the entire image [24] by leveraging pre-trained models with external data (e.g., ImageNet features). The result of DMA-NetImgFu (75.4 %) is obtained by averaging the prediction results of DMA-Net and the fine tuned Alexnet [17]. It is interesting that, though [24] incorporated transformed entire images to represent global information, it still fall behind the performance of our proposed A-Lamp (82.5 %). Such results further validate the effectiveness of our proposed layout-aware subnet. The layout-aware approach slightly boosts the performance of Ours-MP-Net, and significantly performs better than the other state-of-the-art approaches. Such results show that both the holistic layout information and fine-grained information are useful for image aesthetics categorization, and the proposed adaptive Multi-Patch selection approach captures the fine-grained information in compensate to the global view of images.



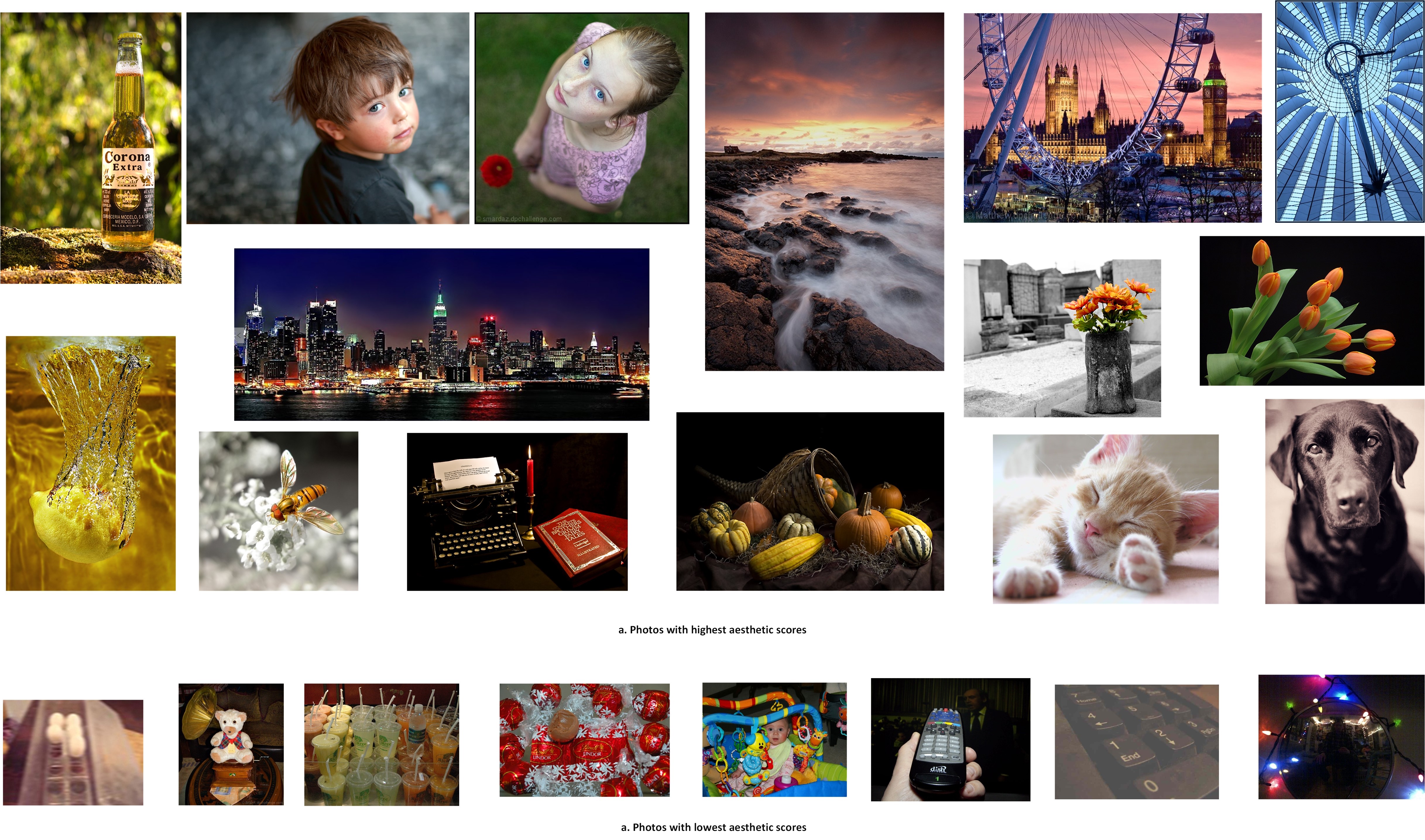
We further examined if our A-Lamp network has learned to respond to the change in image holistic layout and finegrained details. To test this, we random collect 20 highquality images from the AVA dataset. We generate a down sampled version and a wrapped version from the original image. As seen in Fig. 5.1.2, the down-sampled version remains the same aspect ratio (i.e. the layout has not be changed), while half of the original dimension. The wrapped version is generated by scaling along the longer edge to make it square. From the predicted aesthetic score we can notice that, our A-Lamp net give higher score than both the transformed versions. Fig. 5.1.2 shows examples used in the study and their transformed versions, along with our A-Lamp predicted posteriors. The result shows that our A-Lamp is able to reliably respond to the change of image layout and fine-grained details caused by transformation. In addition, we can notice that when the image content is more semantic, it will be sensitive to holistic layout, as we can see, the wrapped version of the portrait photo is much lower than the original and even down-sampled one. It is interesting that the wrapped version for the second photo seems not so bad, while the down-sampled version falls a lot due to much detail loss. To further investigate the effectiveness our A-Lamp networks adaption for content-based image aesthetics, we performed content-based photo aesthetic analysis in the next section 5.2.

## Content-based photo aesthetic analysis

We took the eight most popular semantic tags, i.e. portrait, animal, still-life, food-drink, architecture, floral, cityscape and landscape, as used in [29]. We used the same testing image collection with [23], roughly 2.5K for testing in each of the categories. In each of the eight categories, we systematically compared Ours-MP-Net and A-Lamp net with baseline approach [29] (denoted by AVA) and the stateof-the-art approach [23]. Specifically, SCNNc and SCNNw denote the single-column CNN in [23] that takes centercropping and warpping as inputs, respectively. DCNN denotes the double-column CNN in [23]. As presented in Fig. 5.1.2, the proposed networking training approach significantly outperforms the state-of-the-art in most of the categories, where ”floral” and ”architecture” show much improvements. We find that, photos belonging to these two categories often show complicated texture details, which can can be seen in Fig. 5.1.2. The proposed adaptive MultiPatch subnet remains the fine-grained details, thus turns out much better performance. We also find that A-Lamp show much better performance than Ours-MP-Net in ”portrait” and ”animal”. This result indicates that once an image is associated with an obvious semantic meaning, then the global view is more important than the local view in terms of assessing image aesthetics. Fig.5.1.2 shows some examples of the test images that are considered of the highest and lowest aesthetics values by our A-Lamp. Here we picked good photos from all of the eight categories.

# Conclusion

This paper presents an Adaptive Layout-Aware Multi-

Patch Convolutional Neural Network (A-Lamp CNN) architecture for photo aesthetic assessment. This novel scheme is able to accept arbitrary sized images, and deal with both fined grained details and holistic image layout simultaneously. To support A-Lamp training on these hybrid inputs, we extend the method by developing a dedicated double-subnet neural network structure, i.e. a Multi-Patch subnet and a Layout-Aware subnet. We further construct an aggregation layer to effectively combine the hybrid features from these two subnets. Our experiments on the large-scale AVA benchmark show that our A-Lamp CNN can significantly improve the state of the art in photo aesthetics assessment. ————————————————————— ——–

# References

1. L. Anne Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
2. S. Bhattacharya, R. Sukthankar, and M. Shah. A framework for photo-quality assessment and enhancement based on visual aesthetics. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM ’10, pages 271–280, New York, NY, USA, 2010. ACM.
3. S. Bianco, L. Celona, P. Napoletano, and R. Schettini. On the use of deep learning for blind image quality assessment. *CoRR*, abs/1602.05531, 2016.
4. X. Cao, X. Wei, X. Guo, Y. Han, and J. Tang. Augmented image retrieval using multi-order object layout with attributes. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM ’14, pages 1093–1096, New York, NY, USA, 2014. ACM.
5. D. Cohen-Or, O. Sorkine, R. Gal, T. Leyvand, and Y.-Q. Xu. Color harmonization. In *ACM SIGGRAPH 2006 Papers*, SIGGRAPH ’06, pages 624–630, New York, NY, USA, 2006. ACM.
6. R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *Proceedings of the 9th European Conference on Computer Vision - Volume Part III*, ECCV’06, pages 288–301, Berlin, Heidelberg, 2006. Springer-Verlag.
7. S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1657–1664, June 2011.
8. P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *Int. J. Comput. Vision*, 59(2):167–181, Sept. 2004.
9. L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
10. K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. volume abs/1406.4729, 2014.
11. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
12. Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM ’14, pages 675–678, New York, NY, USA, 2014. ACM.
13. S. Jones and L. Shao. A multigraph representation for improved unsupervised/semi-supervised learning of human actions. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 820–826, June 2014.
14. L. Kang, P. Ye, Y. Li, and D. Doermann. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR ’14, pages 1733–1740, Washington, DC, USA, 2014. IEEE Computer Society.
15. S. Karayev, A. Hertzmann, H. Winnemoeller, A. Agarwala, and T. Darrell. Recognizing image style. *CoRR*, abs/1311.3715, 2013.
16. Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1*, CVPR ’06, pages 419–426, Washington, DC, USA, 2006. IEEE Computer Society.
17. A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.
18. G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891– 2903, Dec 2013.
19. T. Lan, W. Yang, Y. Wang, and G. Mori. Image retrieval with structured object queries using latent ranking svm. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part VI*, ECCV’12, pages 129–142, Berlin, Heidelberg, 2012. Springer-Verlag.
20. L. Liu, R. Chen, L. Wolf, and D. Cohen-Or. Optimizing Photo Composition. *Computer Graphics Forum*, 2010.
21. S. Liu, X. Qi, J. Shi, H. Zhang, and J. Jia. Multi-scale patch aggregation (mpa) for simultaneous detection and segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
22. X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang. Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM ’14, pages 457–466, New York, NY, USA, 2014. ACM.
23. X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang. Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM ’14, pages 457–466, New York, NY, USA, 2014. ACM.
24. X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang. Deep multipatch aggregation network for image style, aesthetics, and quality estimation. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV ’15, pages 990–998, Washington, DC, USA, 2015. IEEE Computer Society.
25. Y. Lu, T. Wu, and S.-C. Zhu. Online object tracking, learning, and parsing with and-or graphs. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR ’14, pages 3462–3469, Washington, DC, USA, 2014. IEEE Computer Society.
26. Y. Luo and X. Tang. Photo and video quality evaluation: Focusing on the subject. In *Proceedings of the 10th European Conference on Computer Vision: Part III*, ECCV ’08, pages 386–399, Berlin, Heidelberg, 2008. Springer-Verlag.
27. L. Mai, H. Jin, and F. Liu. Composition-preserving deep photo aesthetics assessment. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
28. L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV ’11, pages 1784– 1791, Washington, DC, USA, 2011. IEEE Computer Society.
29. N. Murray, L. Marchesotti, and F. Perronnin. Ava: A largescale database for aesthetic visual analysis. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2408–2415. IEEE, 2012.
30. P. Obrador, L. Schmidt-Hackenberg, and N. Oliver. The role of image composition in image aesthetics. In *2010 IEEE International Conference on Image Processing*, pages 3185– 3188, Sept 2010.
31. J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, and N. E. O’Connor. Shallow and deep convolutional networks for saliency prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
32. J. Park, J. Y. Lee, Y. W. Tai, and I. S. Kweon. Modeling photo composition and its application to photo re-arrangement. In *2012 19th IEEE International Conference on Image Processing*, pages 2741–2744, Sept 2012.
33. F. Perronnin, J. Sanchez, and T. Mensink.´ Improving the fisher kernel for large-scale image classification. In *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ECCV’10, pages 143–156, Berlin, Heidelberg, 2010. Springer-Verlag.
34. S. Reed, Z. Akata, H. Lee, and B. Schiele. Learning deep representations of fine-grained visual descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
35. J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888– 905, Aug. 2000.
36. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.
37. H.-H. Su, T.-W. Chen, C.-C. Kao, W. H. Hsu, and S.Y. Chien. Scenic photo quality assessment with bag of aesthetics-preserving features. In *Proceedings of the 19th ACM International Conference on Multimedia*, MM ’11, pages 1213–1216, New York, NY, USA, 2011. ACM.
38. H. Tang, N. Joshi, and A. Kapoor. Blind image quality assessment using semi-supervised rectifier networks. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR ’14, pages 2877–2884, Washington, DC, USA, 2014. IEEE Computer Society.
39. X. Tang, W. Luo, and X. Wang. Content-based photo quality assessment. *Trans. Multi.*, 15(8):1930–1943, Dec. 2013.
40. C.-L. Wen and T.-L. Chia. The fuzzy approach for classification of the photo composition. In *2012 International Conference on Machine Learning and Cybernetics*, volume 4, pages 1447–1453, July 2012.
41. H. Xu, J. Wang, X.-S. Hua, and S. Li. Image search by concept map. In *Proceedings of the 33rd International ACM SI-*

*GIR Conference on Research and Development in Information Retrieval*, SIGIR ’10, pages 275–282, New York, NY, USA, 2010. ACM.

1. N. Xu, B. Price, S. Cohen, J. Yang, and T. S. Huang. Deep interactive object selection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
2. C. Yang, L. Zhang, H. Lu, X. Ruan, and M. H. Yang.

Saliency detection via graph-based manifold ranking. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3166–3173, June 2013.

1. L. Yao, P. Suryanarayan, M. Qiao, J. Z. Wang, and J. Li. Oscar: On-site composition and aesthetics feedback through exemplars for photographers. *Int. J. Comput. Vision*, 96(3):353–383, Feb. 2012.
2. J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech.˘ Unconstrained salient object detection via proposal subset optimization. In *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2016.
3. Z. Zhou, S. He, J. Li, and J. Z. Wang. Modeling perspective effects in photographic composition. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM ’15, pages 301–310, New York, NY, USA, 2015. ACM.