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, warping, 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

Problems of image aesthetics assessment have drawn numerous research attentions with the goal of endow computers with the capability of perceiving aesthetics and visual quality as human vision systems. Potential usage for this task could be foreseen towards wide contemporary applications from intelligent computer systems to real-time, mobile applications.

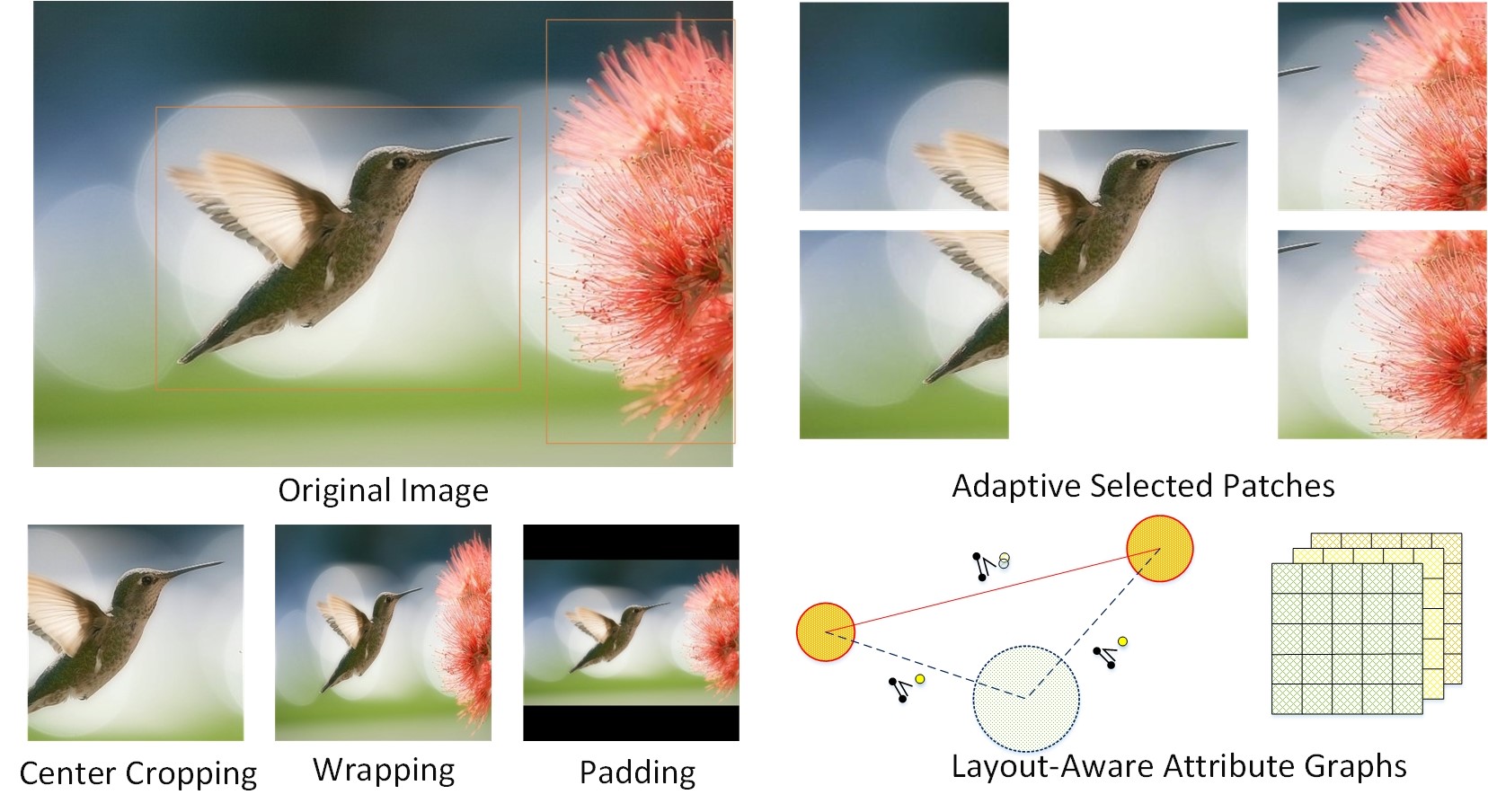


Figure 1. Illustration fo image transformation.

However, assessing photo aesthetics is challenging. Early methods [6, 16] manually design various hand-craft aesthetics features to approximate a number of photographic and psychological aesthetics rules, including lowlevel features [26, 2] (distribution of edges, color histograms and light contrast, etc.), as well as high level features [39, 7, 37, 5] (composition principles, e.g. ”Rule of Thirds”, ”Visual Balance” and ”Golden Ratio”, Low-ofDepth, color harmony, photo content and scene categories, etc.). Although these handcraft features have shown encouraging results. Manually design effective aesthetics features is still a challenging task because even experienced photographers use very abstract terms to describe high quality photos. Other approaches have 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. Though obtaining promising performance, the image representation provided by those generic features may not be optimal for the task of photo aesthetics, as they are designed to represent natural images in general, not specifically for aesthetics assessment.

Because of these limitations, many researchers are attempting to use deep learning methods to extract effective aesthetics features [24, 22, 27, 38, 15]. Although these deep CNN methods have shown promising results, the performance 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, warping, or padding. These operations often alter image composition, reduce image resolution, or cause image distortion, and thus impair the aesthetics of the original images because of potential loss of fine grained details and holistic image layout. However, such fine-grained details and image layout have been shown highly useful in many applications such as 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 as aesthetic attributes in one patch may not well represent the holistic information in the entire image (the flower is eliminated and only the bird is remained). Uniformly warping reduces original image resolution and distorts the salient object thus compromises the detail clarity of the important regions. The artificial boundaries between the original image and the padding area could possibly confuse the neural network. Finally, training from such transformed images will likely make the data more ambiguous and thus compromise the ability of the network to learn effective discriminative features.

Some works address the fixed-size restriction by training images in a few different scales to mimic varied input sizes [10, 27]. However, they still learn from transformed images, which may result in loss of fine grained details and distortion of image layout. To support learning from finegrained details, [24] proposed a deep multi-patch aggregation network architecture (DMA-Net) to take multiple random cropped patches as input. This network shows promising results. However, these orderless bag of patches cannot represent image layout, which result in the global information missing. Moreover, the random cropping strategy requires a large number of training epochs, which lead to very low efficiency.

To resolve the above mentioned issues, we present an

Adaptive Layout-Aware Multi-Patch Convolutional Neural Network (A-Lamp CNN) architecture for photo aesthetic assessment. This dedicated CNN can accept arbitrary image sizes, and deal with both the fine-grained details as well as image layout simultaneously. 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 conventional random cropping methods, we propose an adaptive multi-patch selection scheme to enhance the training efficiency and achieve significant performance improvement. More importantly, we also learn from the holistic image composition by representing images local and global layout leveraging attribute graphs. Finally, aggregation layer is adopted to effectively combine the hybrid features from 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. 1.1. Challenges and Contribution

Learning from both fine-grained details and image layout is challenging. First, the detail information locates in original, relatively high resolution images. Training deep networks with large-size input dimensions requires much longer training time and a significantly larger network structure, training dataset, and hardware memory. To implement fine-grained details learning in practical, we formulate the problem by representing an input image with a small set of carefully cropped patches and associating the set with the image’s label. An aggregation structure leverage statistical functions is adopted to incorporate the multiple patch instances. More importantly, to enhance the training efficiency, we propose an adaptive multi-patch selection strategy instead of previous random cropping method [24]. The central idea is to maximize the efficient input information. We realize that by dedicatedly selecting the patches that play important role in affecting images’ aesthetics. Experimental evaluation demonstrates that, using much less training epochs, our A-Lamp outperformed the performance of

[24].

Second, effectively describing specific image layout and incorporating it into the deep CNN is not straightforward. Previous image layout descriptors are dominantly based on some simple photography composition principles, such as visual balance, rule of thirds, golden ratio, etc. However, these general models cannot represent local and global image layout specifically. To incorporate global information into CNN, [23] used transformed images (warping and center-cropping) to represent the global view. However, such transformation often alter the original image composition or cause distortion. Therefore, this simple strategy is far from enough (We will show the performance evaluation in section 5).

In this paper, 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 is described using object-specific local attributes, and the overall scene with global attributes, thereby capturing both local and global descriptions of the image specifically. The experimental evaluation shows that modeling image layout by constructing attribute-graphs results in improved performance over existing approaches.

Based on the above descriptions, our main contribution can be summarized into three-fold:

* We introduce novel neural network architecture to support learning from original images without considering the image size restriction.
* In particular, we propose two novel subnets to support learning from fine-grained details and holistic image layout. Moreover, aggregation strategies are developed to effectively combine these hybrid information.
* To enhance the training efficiency, we propose an adaptive patch selection strategy, which demonstrate significant improvement over the state of the art.

# Related Work

## Deep Convolutional Neural Networks

Recently, deep learning methods have shown great success in various computer vision tasks, including conventional tasks (e.g. object recognition [42], object detection [10, 21], and image classification [34, 11], etc.) and higher level tasks (e.g. image captioning [1], saliency detection [31], style recognition [9, 15] and photo aesthetics assessment [22, 24, 38, 27, 14], etc.). Most of the existing methods transform input images via cropping, scaling, and padding to accommodate the deep neural network architecture requirement, which compromise the network performance as discussed in section 1.

Recently, [10] and [27] construct adaptive spatial pooling layers trying to alleviate the fixed-size restriction. Theoretically, these network structures can be trained with standard back-propagation, regardless of the input image size. But in practice, the GPU implementations are preferably run on fixed input size. Thus they mimic the varied input sizes by using multiple fixed-size inputs which are scaled from original images. It is apparently far from arbitrary size input. Moreover, they still learn from transformed images, which compensate the network performance, as discussed in section 1.

Other methods propose dedicated network architectures. [22] developed a double-column deep convolutional neural network to support heterogeneous inputs, i.e., global and local views. The global view is represented by padded or warped image, and 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 simultaneously take multiple random cropped patches as input. This network shows promising results. However, these orderless bag of patches cannot represent image layout, which result in the heuristic information missing. Moreover, to ensure that most of the information will be fed into the network, they randomly select 50 groups of patches for each of the image, and train them for 50 epochs, which turns out very low training efficiency.

## Image Layout Representation

To represent holistic image layout, previous works [20, 30, 32, 40, 44] are dominantly model image composition by approximating some simple traditional photography composition guidelines, such as visual balance, rule of thirds, golden ratio, diagonal dominance, etc. However, these simple guidance-based descriptors are not optimal for the task of photo aesthetics. Because they cannot represent complicated image layout specifically.

Attribute-graph, which has long been used by the vision community ro represent structured groups of objects [8, 25, 13, 35], shows promising results in representing complicated image layout. [19] considers the spatial relationship between a pair of objects, while they do not account for the overall geometrical layout of all the objects and the object characteristics. [41] maintains spatial relationships but do not consider background information and object attributes. [18] considers both objects and their interrelations, but do not model the background holistically. [4] perform image ranking by constructing triangular object structures with attribute features. However, they fail to take into account other important aspects such as the global scene context.

To resolve the above mentioned issues, this paper designs a dedicated CNN architecture (A-Lamp). It can accept arbitrary images with its native size. Training and testing are performed under considering both fine-grained details and image layout, thus preserving the quality of the original images. The design of our proposed A-Lamp CNN is inspired by the success of fine-grained detail learning using multi-patch strategy [24, 21], and holistic layout representation by attribute graph. Like DMA-Net in [24], our method also crops multiple patches from original images to preserve fine-grained details. Compared to DMA-Net, our method has two main differences. First, instead of cropping patches randomly, we propose an adaptive patch selection strategy. Second, unlike the DMA-Net that just focus on the fine-grained details, our A-Lamp CNN incorporates the holistic layout. The experimental results demonstrate great enhancement with regarding to efficiency and accuracy over the DMA-Net.

# Adaptive Layout-Aware Multi-Patch CNN

The architecture of the proposed A-Lamp is shown in Fig. 3. Given an arbitrary sized image, multiple patches will be carefully selected by the *Adaptive Patch Selection* module, and fed into the *Multi-Patch subnet*. A statistic aggregation layer is followed to effectively combine the extracted features from these multiple channels. At the same time, a trained CNN is adopted to detect salient objects in the image. The local and global layout of the input image are further represented by an *Attribute-Graph*. At the end, a

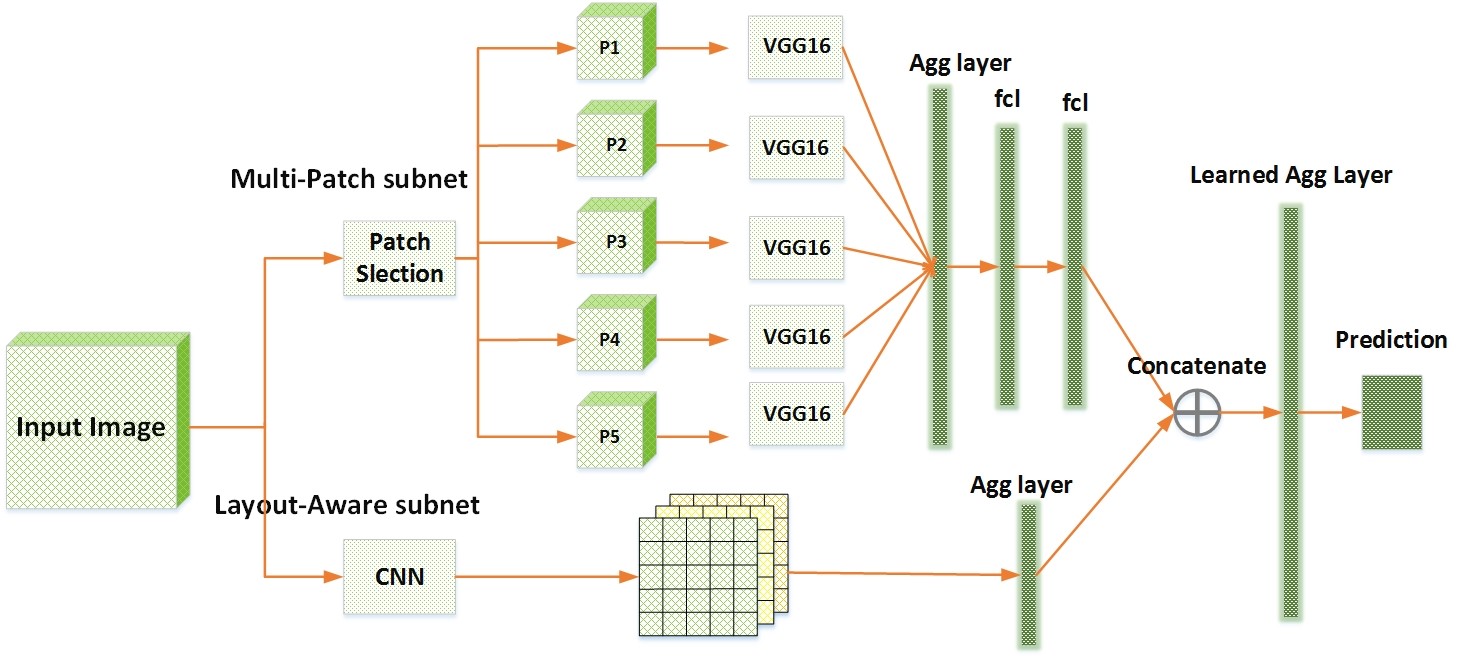


Figure 2. The architecture of the A-Lamp

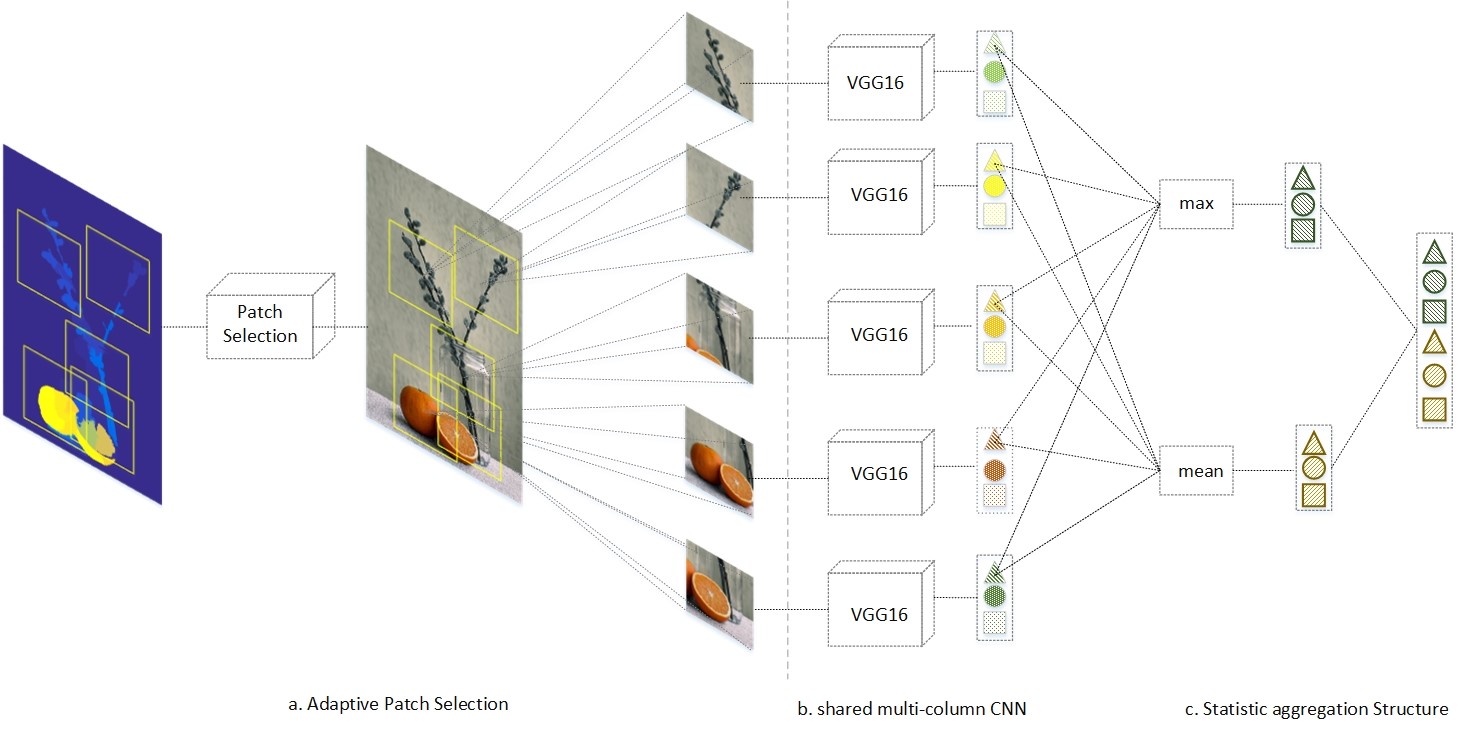


Figure 3. The architecture of Multi-Patch subnet. The Multi-Patch subnet consists of three parts: an adaptive patch selection module (Fig. 3.1 a), a set of paralleled CNNs that are used for extracting deep features from each of the patch (Fig. 3.1 b), and an orderless aggregation structure which combines the extracted deep features from the multi-column CNNs jointly (Fig. 3.1 c).

learning-based aggregation layer is utilized to incorporate the hybrid features from the two subnets and finally produce the aesthetic prediction. More details will be illustrated in this section.

## Multi-Patch subnet

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. The architecture of proposed *Multi-Patch subnet* is shown in Fig.3.1 and more details will be explained in this section.

### Adaptive Patch Selection

Different from the random-cropping method in [24], we aim to carefully select the most discriminative and informative patches to enhance the training efficiency. To realize that, we studied professional photography rules and human visual principles. It has been observed that, human visual attention does not distribute evenly within an image. That means some regions play more important roles than other regions when people viewing photos. In addition, holistic analysis is critical for evaluating an image’s aesthetics. It has been shown that focusing on the subjects is often not enough for overall aesthetic assessment. Motivated by these observations, several criteria have been developed to perform patch selection:

1. Saliency Map. The task of saliency detection is to identify the most important and informative part of a scene. Saliency map models human visual attention, and is capable of highlighting visually significant region. Therefore, it is natural to adopt saliency map for selecting regions that human usually pay more attention to.
2. Pattern Diversity. In addition to saliency map, we also encourage diversification within a set of patches. Different from conventional computer vision tasks, such as image classification and object recognition, that often focus on the foreground objects, image aesthetics assessment also depends on heavily on holistic analysis of entire scene. Important aesthetic characteristics, e.g. Low-of-Depth, color harmonization, and simplicity, can only be perceived by analyzing both the foreground and background as a whole.
3. Overlapping Constraint. Spatial distance among any patch pair should also be considered to minimize the overlapped ratio of these selected patches.

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

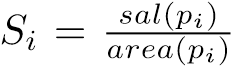
{*c*∗} = *argmaxF* (*S, Dp,Ds*) (1)

*i,j*∈[1*,M*]

*M M M*

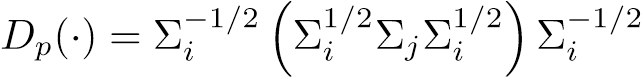
*F*(·) = X*Si* + X*Dp*(N˜*i,*N˜*j*) + X*Ds*(*ci,cj*) (2)

*i*=1 *i*6=*j i*6=*j*

where {*c*∗*m*}*m*∈[1*,M*] is the centers of the optimal set of *M* selected patches.  is the normalized saliency value for each patch *pi*. 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 distributions to represent the pattern of each patch. We model the edge and chrominance distribution of a patch using a multi-variant Gaussian:

*N*˜*m* = {{*Ne*(*µe,*Σ*e*)}*m,*{*Nc*(*µc,*Σ*c*)}*m*}*m*∈[1*,M*] (3) where {*Ne*(*µe,*Σ*e*)}*m* and {*Nc*(*µc,*Σ*c*)}*m* denote the edge distribution and chrominance distribution of patch

*pm*, respectively. P*e* and P*c* is the covariance matrices of edge distribution *Ne* and chrominance distribution *Nc*. Therefore, measuring pattern difference between a pair of patches can be formulated by mapping these distributions to the *Wasserstein Metric space Mm*×*m*, and calculate the 1*stWassersteindistance* between the two distributions *N*˜*i* and *N*˜*j* on this given metric space *M*.

 (4)

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

### Orderless Aggregation Structure

Inspired by [24], we perform the aggregation of the multiple instances to enable the propsoed network learn from multiple patches cropped from a given image. Let be the set of patch features extracted from *nth* image at *lth* layer of the shared CNNs. *bni,l* is a *K* dimensional vector. *Tk* denotes the set of values of the *kth* component of all *bni,l* ∈ h*Blobn*i*l*. For simplicity, we omit image index *n* and layer index *l*, thus *Tk* = {*dik*}*i*∈[1*,M*].

The 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. The outputs of the functions in *U* are concatenated to produce a *Kstat* dimensional feature vectors. Two fully connected layers are followed to implement multi-patch aggregation component. The whole structure can be expressed as a function *f* : {*Blob*} → *Kstat*:

*f*(*Blob*) = *W* × (⊕*Uu*=1 ⊕*Kk*=1 *FAggu* (*Tk*)) (5)

where ⊕ is a vector concatenation operator which produces a column vector, *W* ∈ *Kstat*×*UK* is the parameters of the fully-connected layer. To alleviate the problem of over-fitting, we modified the aggregation function in [24] [[1]](#footnote-1). We adopt two statistical functions that show best performance after conducting extensive experiments. These two functions constitute the vector *U* = {*max, mean*}. 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

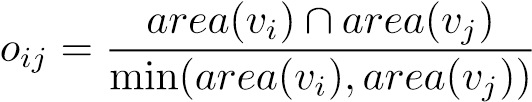
The layout of an image is another critical element that affects the aesthetics. Balanced composition will make an image look more appealing even for an ordinary scene. In this paper, we develop a novel Layout-aware subnet to be combined with the Multi-Patch subnet described early to effectively enhance the performance of the proposed A-Lamp CNN. To properly represent the image layout, we construct an *Attribute-graph*, which is an undirected fully connected graph, by incorporating both local and global image characteristics. As can be seen from Fig. 3.2, the graph nodes characterize the salient objects as well as the overall scene context using node attributes, while the edges connecting nodes capture the topology of the salient objects.

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* represents the nodes and *E* represents the set of edges connecting the nodes. We define two types of attributes in this research:

Local Attributes Φ*local*: Each object present in the image contributes to a graph node resulting in a total of *Nobj* local nodes *Vlocal* = {*v*1*,*· · ·*,vNobj*}. local edges *Elocal* refer to the edges between a pair of local nodes, there will be (*Nobj* − 1)! 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 local features capture the relative arrangement of the objects with respect to each other while the global features define the positioning of all the objects in the image. The attributes can be represented by

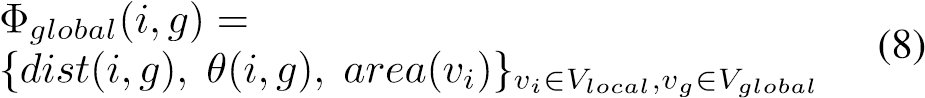
Φ*local*(*i,j*) = {*dist*(*i,j*)*, θ*(*i,j*)*, o*ˆ(*i,j*)}*vi,vj*∈*Vlocal* (6)

where Φ*local*(*i,j*) represents the attribute of a pair of connecting node *vi* and *vj*. *dist*(*i,j*) is the spatial distance between object centroids. *θ*(*i,j*) represents the angle of the graph edge with respect to the horizontal taken in the anticlockwise direction. It indicates the relative spatial organization of the two objects. *o*ˆ(*i,j*) represents the amount of overlap between the bounding boxes of the two objects and is given by

 (7)

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.

Global Attributes Φ*global*: The global node *Vglobal* represents the overall scene including background. The edges connecting local nodes and global node are global edges *Eglobal*, there will be *Nobj* such edges. 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 global attributes are given by



where *dist*(*i,g*) and *θ*(*i,g*) are the magnitude and orientation of the edge connecting the centroid of the object corresponding to node *vi* to the global centroid *cg*. The edges connecting each object to the global node illustrate the placement of that object with respect to the overall object topology.

After that, an aggregation layer is adopted to concatenate the constructed attribute graphs into a feature vector *~ν*, and further combined with the *Multi-Patch subnet*, which can be seen in Fig. 3.[[2]](#footnote-2)

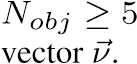
|  |
| --- |
| Figure 4. Architecture of the Layout-Aware subnet |

# Implementation Details

In the implementation, we simplify the computational complexity and release the memory burden in training by first training the Multi-Patch subnet and then combining with the Layout-Aware subnet to fine-tune the overall 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-the-art object-recognition networks that has been successfully applied 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.

# Experimental Results

We systematically evaluate the proposed scheme 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] in which roughly 20,0000 images are used for training and 19,000 images for testing. We also follow the same procedure as previous works 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.

. Therefore, we set *Nobj* =4 to fix the dimension of the feature

|  |  |
| --- | --- |
| Method | Accuracy |
| DMA-Netave | 73.1 % |
| DMA-Netmax | 73.9 % |
| DMA-Netstat | 75.4% |
| DMA-Netfc | 75.4% |
| New-MP-Net | 81.7% |

Table 1. Performance of Multi-Patch subnet

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

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

### Analysis of adaptive Multi-Patch subnet

For a fair comparison, we first perform the training and testing only using the proposed Multi-Patch subnet, and evaluate New-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 is denoted as DMA-Netstat and Fully-Connected Sorting Aggregation Structure as DMA-Netfc.

The experimental results are shown in Table 5.1.1. We can see that, the proposed scheme outperforms all types of DMA-Net architectures. Although existing work DMA-Net [24] randomly cropped 50 groups of patches to train a total of 250 patches for each image and the total training has 50 epochs. The randomness in crop-

ping was not able to effectively capture useful information and may cause the training to be confusing for the network. Instead of random

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

cropping, we adaptively select the most informative and discriminative patches as input, which is the key to achieve substantial performance enhancement. From Fig.1, we can see that, the salient objects, i.e. the bird and the flower, have been selected. Within these patches, the most important information and the fine-grained details are all retained. In addition, the background, i.e. the blue sky and the green ground, have also been selected so that global characteristics, e.g. color harmony, Low-of-Depth, can also be learned. More examples of selected patches are shown in Fig. 5.1.1. We can see that, the proposed adaptive selection strategy not only is effective in selecting the most salient regions (e.g. the human’s eyes and face, the orange flowers, etc.), but also is capable of capturing the pattern diversity (e.g. the green leaf and green beans, the flower and the gray wall).

Furthermore, the proposed adaptive patch selection strategy is also able to enhance the training efficiency. The result of New-MP-Net is obtained by taking 20-30 training epochs, substantially less than 50 epochs reported in [24], while still achieving better performance. Another reason for achieving higher accuracy may also lies in a different CNN architecture we adopted. Shallow CNN with only 4 convolution layers and followed by two fully connected layers is adopted in [24]. In this research, we adopt VGG16 [36] which may also contribute to significant performance improvement.

### A-Lamp CNN Performance

Table 5.1.2 shows the results of the proposed A-Lamp CNN on the AVA dataset [29] for image aesthetics categorization. The AVA dataset 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 recently developed deep CNN schemes outperform the conventional approach.

To examine the effectiveness of the proposed scheme, we compare New-MP-Net and A-Lamp with the baseline methods that take only fixed-size inputs. In particular, we compare with three VGG16-based aesthetics assessment methods; each operates on a different type of transformed input.

VGG16-Crop: The input of the network is obtained by randomly cropping from the original image with a 224×224 cropping window. This cropping window size is fixed as 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 crops with their corresponding aesthetics labels. For each testing image, we follow the existing work in [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-valued pixels.

We can see from Table 5.1.2 that, both the proposed Multi-Patch subnet and the A-Lamp network outperform these fixed-size input VGG nets. Such results confirmed that training network on multiple patches produces better prediction than networks training on a single patch.

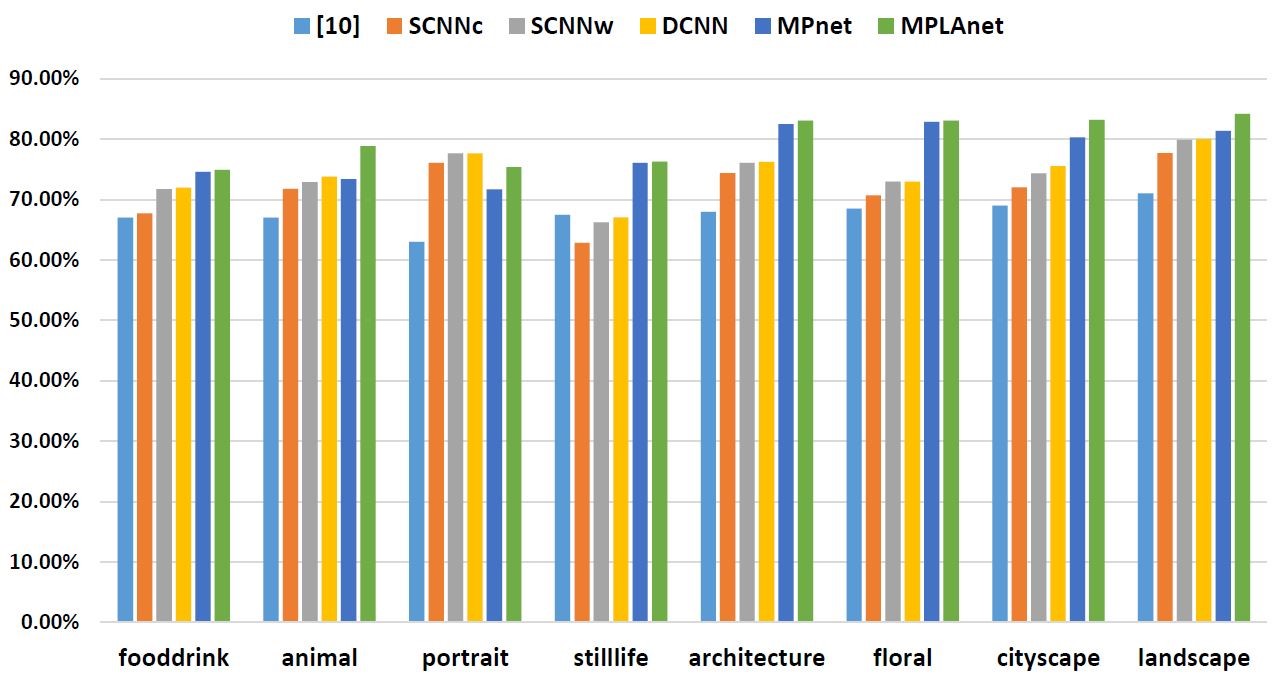
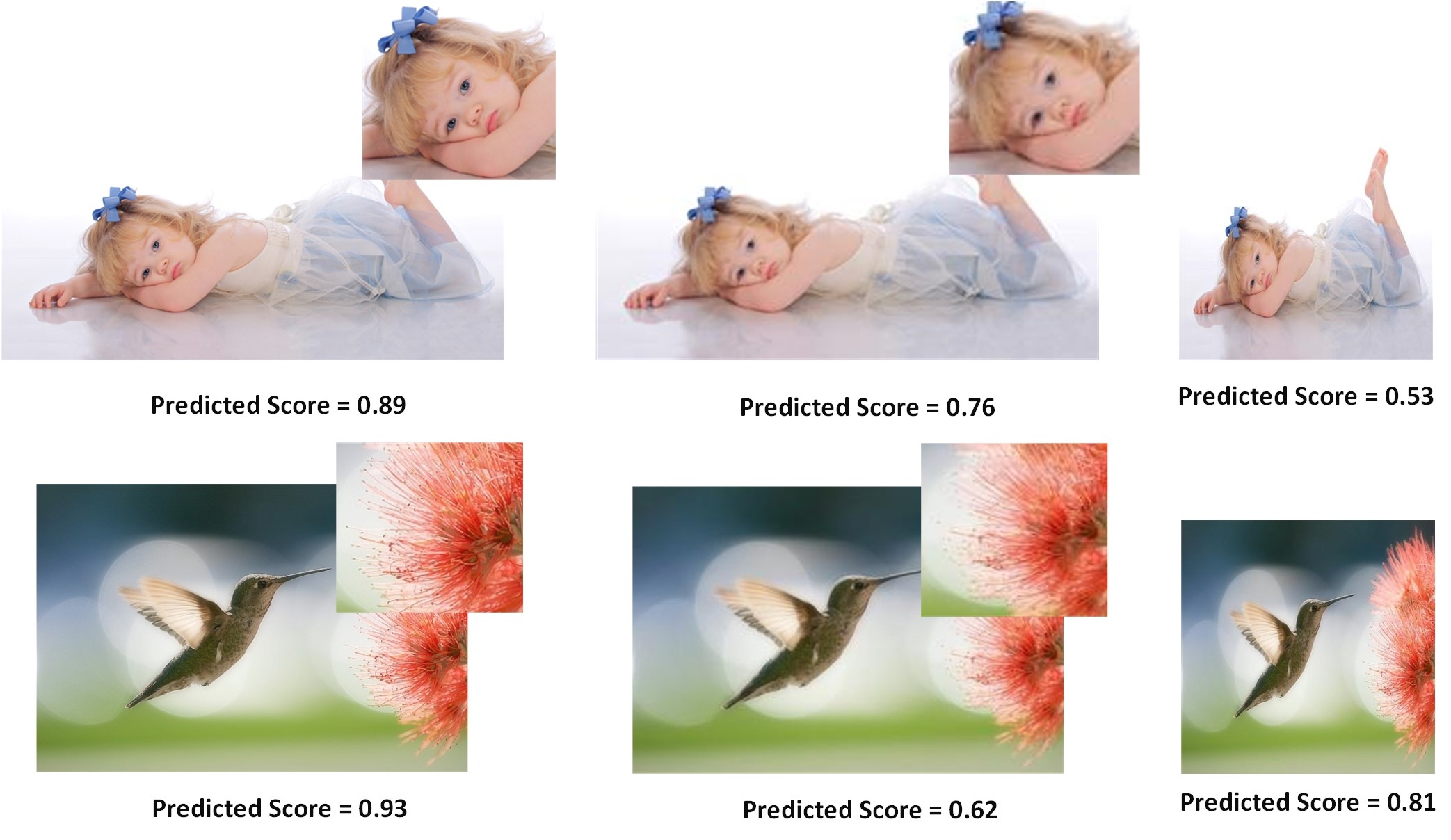
We also compared the proposed scheme with some latest non-fixed size restriction schemes, i.e. SPP-CNN [10] and MNACNN [27]. Different from these schemes that their trainings are from several different level of scaled images, we implement the A-Lamp network to be trained from the original images. The results confirm that learning from original images is essential for aesthetic assessment, as we have discussed earlier. In addition, higher prediction accuracy of the proposed scheme further proves that, the adaptive Multi-Patch strategy we adopted in this research is more efficient than the spatial pooling layers adopted in SPP-CNN and MNA-CNN.

To show the effectiveness of the 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 to produce the final aesthetics prediction. We can see from the results that incorporating scene attributes does not cause noticeable performance improvement.

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

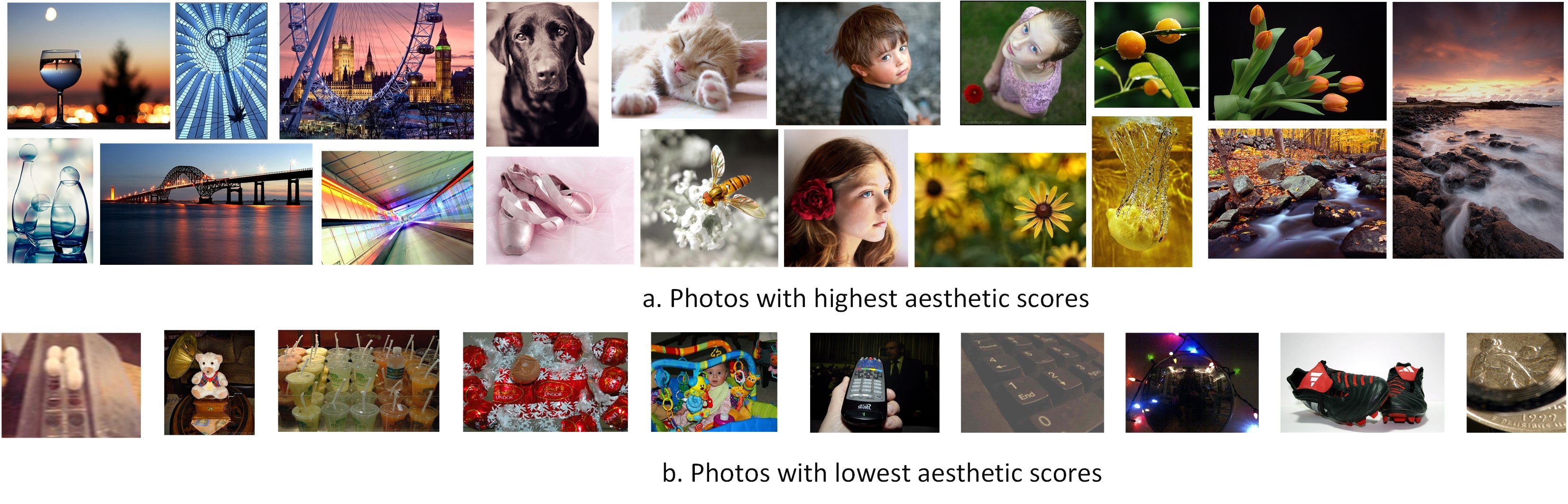
DMA-Net-ImgFu keeps the global view of the entire image [24] by leveraging pre-trained models with external data (e.g., ImageNet features). The result of DMA-NetImgFu (75.4 %) is obtained by averaging the prediction results of DMA-Net and the fine-tuned Alexnet [17]. It is interesting to see that, although the scheme in [24] incorporated transformed entire images to represent global information, its performance still falls behind that of the proposed A-Lamp (82.5 %). This result further validates the effectiveness of the proposed layout-aware subnet. The proposed layout-aware approach boosts the performance of New-MP-Net slightly, but outperforms significantly over the other state-of-the-art approaches. The overall results show that both holistic layout information and fine-grained information are essential for image aesthetics categorization. The proposed adaptive Multi-Patch selection approach captures the fine-grained information to complement the global view information of the images.

We further examined whether or not the proposed A-Lamp network is capable of responding to the changes in image holistic layout and fine grained details. To test this, we random collect 20 high quality images from the AVA dataset. We generate a down sampled version and a warped version from the original image. As shown in Fig. 5.1.2, the down-sampled version keeps the same aspect ratio (i.e. the layout has not be changed) but reduced to one half of the original dimension. The warped version is generated by scaling along the longer edge to make it square. From the predicted aesthetic score we can confirm that, the A-Lamp network produces higher score for the original image than both transformed versions. Fig. 5.1.2 shows examples used in the study and their transformed versions, along with the A-Lamp predicted posteriors. The result shows that the A-Lamp network is able to reliably respond to the change of image layout and fine-grained details caused by the transformations.



In addition, we also notice that when the image content is more semantic, it will be more sensitive to holistic layout. In particular, the warped version of the portrait photo receives much lower score than the original one, or even the down-sampled one. It is interesting to notice that the warped version for the second photo example 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 have performed content-based photo aesthetic study with detailed results presented in the next.

## Content-based photo aesthetic analysis

To carry out content-based photo aesthetic study, we take photos in eight most popular semantic tags used in [29]: portrait, animal, still-life, food-drink, architecture, floral, cityscape and landscape. We used the same testing image collection used in [23], approximately 2.5K for testing in each of the categories. In each of the eight categories, we systematically compared New-MP-Net and A-Lamp network with the baseline approach [29] (denoted by AVA) and the state-of-the-art approach in [23]. Specifically, SCNNc and SCNNw denote the single-column CNN in [23] that takes center cropping and warping, respectively, as inputs. DCNN denotes the double-column CNN in [23]. As shown in Fig. 5.1.2, the proposed network training approach significantly outperforms the state-of-the-art in most of the categories, where ”floral” and ”architecture” show substantial improvements. We find that, photos belonging to these two categories often show complicated texture details, which can be seen in Fig. 5.1.2. The proposed adaptive Multi-Patch subnet keeps the fine-grained details and thus produces much better performance. We also find that A-Lamp networks shows much better performance than New-MP-Net in ”portrait” and ”animal”. These results indicate that once an image is associated with a clear semantic meaning, then the global view is more important than the local views in terms of assessing image aesthetics. Fig.5.1.2 shows some examples of the test images that are considered by the proposed A-Lamp as among the highest and lowest aesthetics values. These photos are selected from all 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 to capture intrinsic aesthetic characteristics from both fined grained details and holistic image layout simultaneously. To support A-Lamp training on these hybrid inputs, we developed a dedicated double-subnet neural network structure, i.e. a Multi-Patch subnet and a Layout-Aware subnet. We then construct an aggregation layer to effectively combine the hybrid features from these two subnets. Extensive experiments on the large-scale AVA benchmark show that this 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.

1. Extensive experiment show that, using four statistical functions, i.e. min, max, mean, median, not result in performance improvement, and even worse due to over-fitting caused by the too large vector dimension [↑](#footnote-ref-1)
2. By statistical study, we find that, the confidence score is very low when [↑](#footnote-ref-2)