This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

# Composition-preserving Deep Photo Aesthetics Assessment

Long Mai
Portland State University
mtlong@cs.pdx.edu

Hailin Jin
Adobe Research
hljin@adobe.com

Feng Liu
Portland State University
fliu@cs.pdx.edu

## Abstract

*Photo aesthetics assessment is challenging. Deep convolutional neural network (ConvNet) methods have recently shown promising results for aesthetics assessment. The performance of these deep ConvNet methods, however, 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, scaling, or padding, which often damages image composition, reduces image resolution, or causes image distortion, thus compromising the aesthetics of the original images. In this paper, we present a composition-preserving deep ConvNet method that directly learns aesthetics features from the original input images without any image transformations. Specifically, our method adds an adaptive spatial pooling layer upon the regular convolution and pooling layers to directly handle input images with original sizes and aspect ratios. To allow for multi-scale feature extraction, we develop the Multi-Net Adaptive Spatial Pooling ConvNet architecture which consists of multiple sub-networks with different adaptive spatial pooling sizes and leverage a scene-based aggregation layer to effectively combine the predictions from multiple sub-networks. Our experiments on the large-scale aesthetics assessment benchmark (AVA [29]) demonstrate that our method can significantly improve the state-of-the-art results in photo aesthetics assessment.*

## 1. Introduction

Subjective photo quality and aesthetics assessment is challenging. Existing photo aesthetics assessment methods extract visual features and then employ various machine learning algorithms to predict photo aesthetic values [2, 5, 7, 12, 13, 17, 25, 27, 28, 31, 32, 42, 48, 51]. Feature extraction is a critical step for aesthetics assessment. Early methods manually design aesthetics features according to people's aesthetics perception and photography rules [2, 5, 7, 17]. Manually designing effective aesthetics features, however, is still a challenging task although these features have shown encouraging results. Other approaches have been developed to leverage more generic image features such as Fisher Vector [28, 35, 36] and bag of visual words [42] to predict photo aesthetics. While obtaining promising performance, the image representation provided by those generic features may not be optimal for photo aes-

(a) Cropping

(b) Scaling

(c) Padding plus scaling

Figure 1: Effect on image transformation on photo aesthetics. (a): Cropping compromises the composition of the originally well-composed image that follows rule of thirds. (b): Scaling distorts the important object. (c): While padding and scaling keeps the original aspect ratio, it sometimes leads to the loss of the image clarity. In this example, the spots on the ladybug is difficult to see in the padding result. The added boundaries between the image and the padding area can also confuse a deep learning algorithm.

thetics as they are designed to represent natural images in general, not specifically for aesthetics assessment.

Deep learning methods, which have shown great success in various computer vision tasks, have recently been used to extract effective aesthetics features [14, 25, 26, 47]. However, applying existing deep learning algorithms, such as deep convolutional network and deep belief network, to aesthetics feature learning is non-trivial. One major challenge is posed by the fixed input size restriction required by the neural networks. This restriction poses a particular challenge for applying a deep neural network algorithm to aesthetics assessment. To meet this restriction, input images need to be transformed via cropping, scaling, or padding before feeding into the neural network. These transformations often compromise the aesthetics of the original images. As illustrated in Figure 1, cropping can sometimes negatively change the image composition, such as turning a well-composed photo in (a) that originally follows rule of thirds into an ill-composed one. Scaling distorts the salient object in (b) and padding plus uniformly scaling reduces the original image resolution and compromises the detail clarity of the important object as shown in (c). Padding also introduces artificial boundaries between the original image and the padding area, and moreover, the locations of these boundaries vary over different images, which could possibly confuse the neural network. Finally, for deep learning, assigning the aesthetics label of an original image to its transformed versions during training will likely make the data more ambiguous and thus compromise the ability of the network to learn good discriminative features.

Existing methods address this fixed-size restriction by designing dedicated convolutional neural network architectures to simultaneously take multiple versions of the transformed images as input [25, 26]. These dedicated networks show promising results; however, they still learn from transformed inputs and it is unclear whether the aesthetics labels of the original images can be transferred to the collection of their transformed versions.

In this paper, we present a deep Multi-Net Adaptive Spatial Pooling Convolutional Neural Network (MNA-CNN) method for photo aesthetics assessment that can directly process the original images without any image transformation. Our method adds an adaptive spatial pooling layer upon regular convolutional and pooling layers. This adaptive spatial pooling layer can handle input images with different sizes and aspect ratios. To allow for multi-scale feature extraction, our deep network architecture consists of multiple sub-networks, each having an adaptive spatial pooling layer with a different pooling size. We further construct a scene-aware aggregation layer to effectively combine the predictions from these multiple sub-networks.

Our MNA-CNN method has a major advantage in that it can directly handle images with their native sizes and aspect ratios, which is critical for aesthetics assessment. Our study shows that by learning from the original images without any transformations, our MNA-CNN network can learn to capture some subtle compositions that are important for aesthetics. Our method is also capable of extracting features at multiple scales and naturally incorporating scene categories for aesthetics assessment. As shown in our experiments, our method can significantly improve the state-of-the-art results for photo aesthetics assessment.

## 2. Related Work

Early methods for image quality assessment measure image quality by detecting and measuring various distortions, including blocking, ringing, mosaic patterns, blur, noise, ghosting, jerkiness, smearing, etc [3, 4, 22, 40, 49, 33, 39, 30, 50, 40]. While they are effective for measuring quality loss due to compression or data loss during transmission, these low-level distortion measurement-based metrics sometimes do not well reflect people's subjective perception of image quality.

Subjective image quality assessment methods have also been developed [2, 5, 7, 12, 13, 17, 25, 27, 28, 31, 42, 48, 51]. Many of these methods represent images using manually crafted features that are carefully designed to approximate a number of photographic and psychological aesthetics rules, such as rule of thirds, visual balance, rule of simplicity, etc [2, 5, 7, 17]. A classifier is then trained using those features to label an input image as low or high quality. Some other approaches directly use generic image features such as Fisher Vector [28, 35, 36] and bag of visual words [42] to predict photo aesthetics.

Recently, deep learning methods have shown great success in various computer vision tasks, such as object recognition, object detection, and image classification [10, 14, 15, 16, 19, 34, 37, 38, 41, 43, 44, 45, 46, 52, 53, 54]. Deep learning methods, such as deep convolutional neural network and deep belief network, have also been applied to photo quality/aesthetics assessment and have shown good results [25, 26, 47]. As most deep neural network architectures require fixed-size inputs, recent methods [25, 26] transform input images via cropping, scaling, and padding, and design dedicated deep network architectures, such as double-column or multi-column networks, to simultaneously take multiple transformed versions as input. Since transformation often affect the aesthetics quality of the original images as discussed in Section 1, this paper designs a dedicated deep Multi-Net Adaptive Spatial Pooling Convolutional Neural Network (MNA-CNN) architecture that can directly process the images with its native size and aspect ratio, thus preserving the quality of the original images.

The design of our MNA-CNN network is inspired by the success of the SPP-Net for visual recognition [11]. Like SPP-Net, our method also constructs an adaptive spatial

pooling layer to allow our network to accept as input images at its original size and aspect ratio. Compared to SPP-Net, our method has two main differences. First, unlike SPP-Net that adopts the training strategy that use multiple fixed-size inputs during training, our method allows arbitrary-size input to be used both in training and testing. Second, instead of using the spatial pyramid pooling layer which concatenates adaptive spatial pooling layers with different sizes together, our method contains multiple sub-networks for different pooling sizes. That allows these sub-networks to learn effective feature detectors for images with different resolutions and aspect ratios and at the same time simplify the fully connected layer which makes the network effective even with a limited amount of training data.

## 3. Composition-preserving Deep Network for Photo Aesthetics Assessment

We first briefly review how the conventional deep convolutional neural network (ConvNet) can be applied to aesthetics assessment and then describe our dedicated deep Multi-Net Adaptive Spatial Pooling ConvNet for photo aesthetics assessment.

**Background**. A deep ConvNet consists of a number of convolutional and pooling layers, followed by some fully connected layers. In this paper, we employ the supervised feature transfer technique that has lead to many successful computer vision applications. Instead of designing and training a new ConvNet from scratch, we reuse a classification ConvNet architecture pre-trained on a large collection of images such as ImageNet [6]. We then modify the top layer of the network to adapt to our aesthetics classification task. In particular, we modify an ImageNet network by turning their 1000-way softmax prediction layer into a single linear unit followed by a sigmoid activation (Figure 2(a)).

The resulted ConvNet represents the mapping function $f_W : I \rightarrow P(Q_I = \text{high}|I)$, where $Q_I$ represents the aesthetics quality of the image $I$. Let $fc_l(I)$ be the output of the last fully connected layer, the sigmoid activation unit models the posterior probability of the input image having a high aesthetics quality as

$$P(Q_I = \text{high}|I) = \frac{1}{1 + e^{-fc_l(I)}} \quad (1)$$

The model is trained on a collection of training examples $S = \{I_n, y_n\}$, where $y_n$ is the binary aesthetics label (high or low) of the image $I_n$. Let $W$ be the set of weights from all the layers of the network. During training, the optimal value of $W$ is determined by minimizing the following binary cross-entropy objective function typically using a stochastic gradient descend algorithm.

$$l(W) = \sum_{n=1}^{N} y_n \log(f_W(I_n)) + (1 - y_n) \log(1 - f_W(I_n)) \quad (2)$$

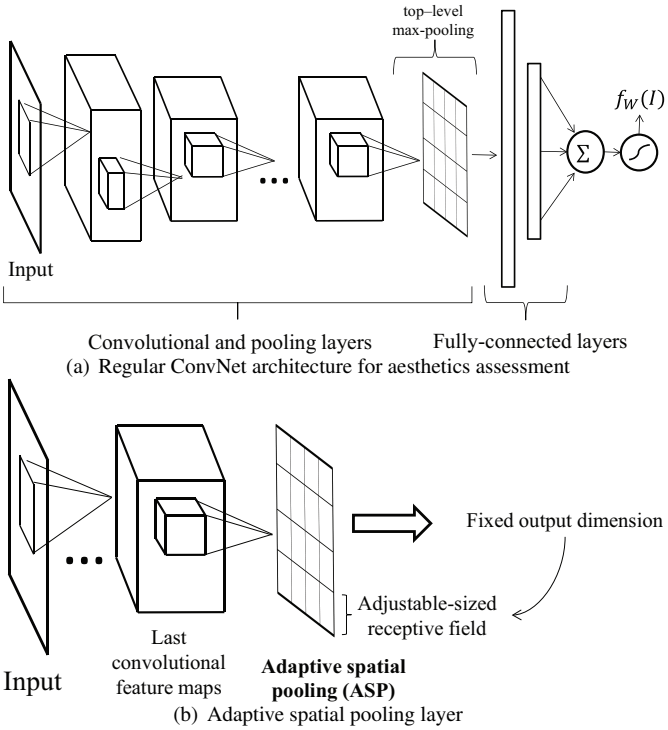



Figure 2: ConvNet architecture for aesthetics assessment and adaptive spatial pooling layer. Similar to the conventional pooling layer, the adaptive spatial pooling layer (ASP) performs the pooling (e.g. max pooling) operator over local image regions. However, instead of fixing the receptive field's size, ASP fixes the output dimension while adjusting the size of the receptive field to handle images with different sizes and aspect ratios.

The major challenge in applying existing ConvNets to photo aesthetics assessment is the fixed-size constraint. Due to this restriction, input images need to be transformed to the pre-defined size before given to the network. As discussed in Section 1, such transformation can sometimes severely affect the ability of the network to learn useful features for aesthetics analysis because transforming the images would often compromise the important factors of the image aesthetics perception such as composition, detail clarity, and/or image content.

### 3.1. Composition-preserving Deep ConvNet

To remove the fixed-size constraint, we employ the adaptive spatial pooling strategy [11] to enable the ConvNet to operate on an image in its original form during both training and testing. Below, we first describe the concept of adaptive spatial pooling and then elaborate how our method incorporates adaptive spatial pooling and develop a Multi-Net Adaptive-Pooling ConvNet that works with images with their original sizes and aspect ratios.

**Adaptive Spatial Pooling.** As discussed in previous work [11, 23], the requirement of fixed-size input imposed by existing ConvNets is due to the last pooling layer of the convolutional structure. The last pooling layer produces the inputs for the subsequent fully-connected layer which demands fixed-size inputs. The main problem is that conventional pooling layers pre-define the size of the local receptive fields (i.e. the local regions on which to pool). The fixed-size receptive field constraint makes the output dimension of a conventional pooling layer depend on its input dimension. As a result, the input image size needs to be fixed in order for the network to generate the input of the dimension required by the fully-connected layers.

Inspired by the spatial pyramid pooling method [11], we relax that problematic fixed-size constraint by employing an alternative pooling strategy: *adaptive spatial pooling*. As illustrated in Figure 2(b), the adaptive spatial pooling layer performs the pooling operator (e.g. max pooling) over local image regions similarly to the conventional pooling layer. However, different from the conventional pooling layers where the size of the receptive field is fixed, the adaptive spatial pooling layer instead fixes the output dimension, and adjust the receptive field size accordingly. This allows the adaptive spatial pooling layer to generate the fixed-size output from input with various sizes. Any existing ConvNet structure can then be modified to accept arbitrary-size input images by replacing the last conventional pooling layer with the adaptive spatial pooling layer.

### 3.1.1 Multi-Net Adaptive-Pooling ConvNet

The importance of multi-scale feature extraction has been emphasized in various computer vision and deep learning research [1, 9, 20, 21]. Using a single pre-defined size for the top pooling layer often restricts the scale at which the lower level features are extracted. The recent SPP-Net method addresses this problem with a spatial pyramid strategy that uses multiple adaptive pooling layers with different sizes and concatenates their outputs [11]. While such a spatial pyramid pooling method allows multi-scale pooling, it restricts all pooling components to share the same lower-level features, which makes it more difficult to learn dedicated features specifically for different pooling layer sizes. More importantly, as the upper fully connected layer is connected to all the pooling layer, it often needs to learn the interactions between them, which makes the learning task more complex and thus requires a large amount of training data. To address these problems, we develop a Multi-Net Adaptive-Pooling method to combine adaptive pooling layers of different sizes.

Our Multi-Net Adaptive-Pooling ConvNet (MNA-CNN) consists of multiple sub-networks, each of which is a copy of the base network with the last pooling layer replaced by an adaptive spatial pooling layer with a specific scale, as illustrated in Figure 3. All sub-networks share the same input image and their outputs are combined with the average operator to obtain the overall prediction. During training, we train each sub-network separately instead of training the whole architecture at the same time so as to minimize the correlation among those networks, which has been shown to improve the ensemble performance [8, 55].

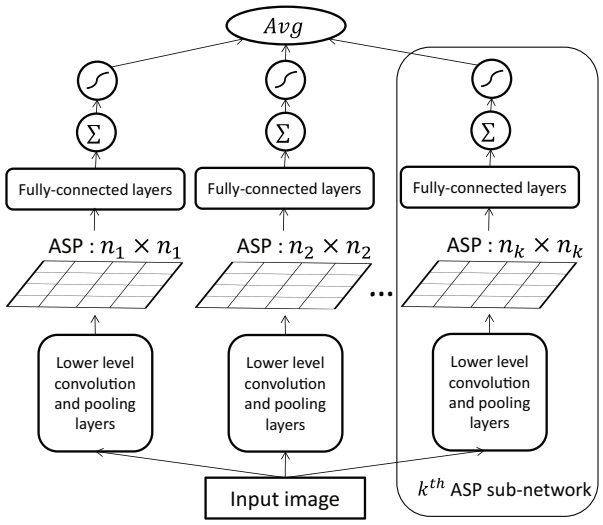



Figure 3: Multi-Net Adaptive-Pooling ConvNet (MNA-CNN). Our MNA-CNN network contains multiple sub-networks, each being a copy of the base network with the last pooling layer replaced by an adaptive spatial pooling layer (ASP) with a specific scale. All sub-networks share the same input image and their outputs are combined with the average operator to obtain the overall prediction.

### 3.1.2 Scene-Aware Multi-Net Aggregation

Our MNA-CNN method averages the prediction results of multiple sub-networks as the final output. While taking the average can leverage the complementary among the sub-networks to improve the overall prediction results as shown in Section 4.2, it treats all sub-networks equally regardless of the image content. Previous research has shown that taking the scene category of the image into account can improve the aesthetics prediction accuracy [48]. Accordingly, we enhance our MNA-CNN method with a learning-based aggregation component to combine the results of sub-networks in a scene-aware manner.

Specifically, we augment our MNA-CNN network with a state-of-the-art scene-categorization deep network [54]. We 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. Figure 4 shows our scene-aware MNA-CNN network that implements the scene-aware aggregation component using a fully-connected layer with 50 neurons.

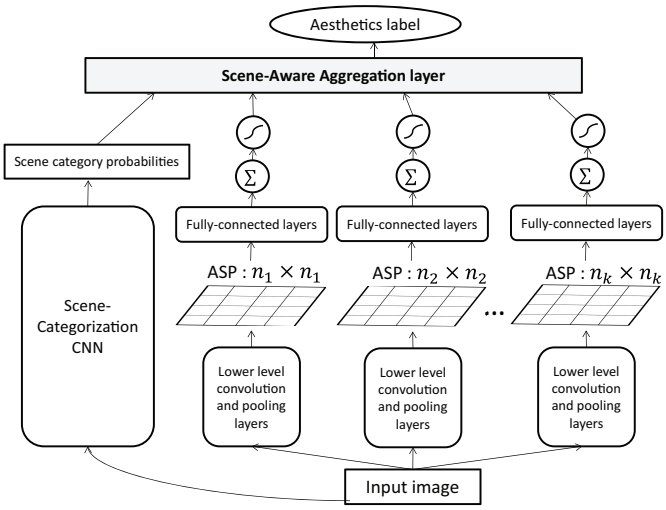



Figure 4: Scene-Aware Multi-Net Aggregation. We augment our network with a scene-categorization deep network. The top-level classifier takes the sub-network predictions and the image scene-categorization posteriors as feature vectors and produce the final aesthetics classification.

Conceptually the scene-aware aggregation layer can be trained end-to-end along with the sub-networks. In our implementation, we simplify the computational complexity in both training and testing by first training the MNA-CNN sub-networks and then training the aggregation on the validation data with the sub-networks fixed.

### 3.2. Implementation Details

Our implementation of the MNA-CNN architecture uses $k = 5$ adaptive pooling layer sizes. Specifically, we implement our sub-networks with the adaptive pooling layer sizes of $12 \times 12$, $9 \times 9$, $6 \times 6$, $4 \times 4$, and $2 \times 2$, respectively. All the network training and testing are done using the Torch deep learning package[1]. The networks are trained with the standard back-propagation algorithm. During training, we fix the learning rate at 0.05 without learning rate shrinkage.

We use the VGG network (VGG-Net) [41] pre-trained on the ImageNet dataset as our base network architecture for supervised feature transfer. VGG-Net is one of the state-of-the-art object-recognition networks that has been adopted with great success to many different computer vision problems. Our experiments show that combining VGG-Net architecture with our MNA-CNN method can significantly improve the aesthetics assessment accuracy compared to the state-of-the-art photo aesthetics methods. The pre-trained VGG network model used in our implementation is obtained from the BVLC CAFFE model zoo[2].

For the scene-categorization ConvNet component, we use the Places205-GoogLeNet[3] ([54]) which was pre-trained on 205 scene categories of Places Database with 2.5 million images. The original scene categorization network Places205-GoogLeNet is trained to recognize 205 scene categories. That is larger than the number of sub-networks in our MNA-CNN architecture. To avoid the feature vectors to the aggregation layer being dominated by the scene categorization prediction, we fine-tune the Places205-GoogLeNet to predict only seven scene categories: *human, plant, architecture, landscape, static, animal, and night*. This set of categories was suggested in [48] and [29] as being very related to aesthetics perception. We obtain the training data for fine-tuning by downloading the images from the web with the keywords as the seven corresponding category names. Specifically, we downloaded about 10,000 images for each category from Flickr[4].

| Method | $\delta = 1$ | $\delta = 0$ |
|---|---|---|
| Murray *et al.* [29] | 67.0% | 66.7% |
| Lu *et al.* [24] | 74.2% | 75.4% |
| Lu *et al.* [26] | N/A* | 75.4% |
| MNA-CNN | 76.1% | 77.1% |
| MNA-CNN-Scene | **76.5%** | **77.4%** |

Table 1: Comparison with the state of the art methods. *This result is not reported in the original paper [26].

## 4. Experiments

We experimented with our method on the AVA benchmark [29], which, to our best knowledge, is the largest publicly available aesthetics assessment benchmark. The AVA benchmark 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 [24, 25, 26, 29]. That is, we allocate 235,599 images for training and 19,930 images for testing.

We follow the same procedure as the previous work [24, 25, 26, 29] to assign a binary aesthetics label to each image in the benchmark. Specifically, images with mean ratings smaller than $5 - \delta$ are labeled as low quality and those with mean ratings larger than or equal to $5 + \delta$ are labeled as high quality. Images in the middle range $[5 - \delta, 5 + \delta]$ are considered ambiguous and discarded. Two different values of $\delta$: $\delta = 0$ and $\delta = 1$ are used to generate the ground truth labels for the training images and $\delta = 0$ is used for all testing images, as suggested in [29].

[1] http://torch.ch/
[2] https://github.com/BVLC/caffe/wiki/Model-Zoo
[3] http://places.csail.mit.edu/downloadCNN.html
[4] https://www.flickr.com/

(a) Photos of highest predicted aesthetics values

(b) Photos of lowest predicted aesthetics values

Figure 5: Aesthetics quality prediction. The top and the bottom show the images with the highest predicted aesthetics values and those with the lowest predicted aesthetics values in the testing dataset, respectively.

### 4.1. Comparison with the State of the Art

We compare our methods MNA-CNN and scene-aware MNA-CNN to the state-of-the-art methods [24, 26, 29]. Here [29] provides the state-of-the-art result for methods that use manually designed features and/or generic image features for aesthetics assessment. [24, 26] are the very recent methods that also design a dedicated deep ConvNet for aesthetics assessment. The results of these methods are obtained from their papers. As shown in Table 1, both our methods outperform the state-of-the-art methods for aesthetics assessment. The comparisons, especially those between our methods and the existing deep ConvNet methods [24, 26], show that preserving the original image size and aspect ratio can most likely lead to improved aesthetics assessment performance. Figure 5 shows some examples of the test images that are considered of the highest and lowest aesthetics values by our scene-aware MNA-CNN method.

### 4.2. Effectiveness of Adaptive Spatial Pooling

To examine the effectiveness of the adaptive spatial pooling layers in our composition-preserving deep aesthetics methods, we compare our methods to the baseline methods with fixed-size inputs. In particular, we experiment with three VGG-Net based aesthetics assessment methods, each operating on a different type of transformed input.

**VGG-Crop:** The input of the network is obtained by randomly cropping the original input image with a $224 \times 224$ cropping window. This cropping window size is the fixed size required by the VGG-Net 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 [25] to predict the aesthetics quality for 50 random crops obtained from the image and take their average as the final prediction result.

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

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

We also experiment with an alternative composition-aware method that makes use of the spatial pyramid pooling layer (SPP-CNN) [11] to handle images with different sizes and aspect ratios. We note that different from the method presented in [11] that only allows an arbitrary-size input during testing, we implement an SPP-CNN network to allow arbitrary-size inputs to be used both during training and testing, which is critical for aesthetics assessment, as discussed in Section 1.

Table 2 compares the performance of the above deep network aesthetics assessment methods in terms of three metrics: classification accuracy, F-measure, and area under the ROC curve (AUC score). The accuracy and F-measure of a network are obtained by binarizing the network's outputs with the threshold value of 0.5 and comparing the results to

| Method | Accuracy | F-measure | AUC score |
|---|---|---|---|
| VGG-Crop | 71.2% | 0.83 | 0.66 |
| VGG-Scale | 73.8% | 0.83 | 0.74 |
| VGG-Pad | 72.9% | 0.83 | 0.73 |
| SPP-CNN | 76.0% | 0.84 | 0.77 |
| MNA-CNN | **77.1%** | **0.85** | **0.79** |

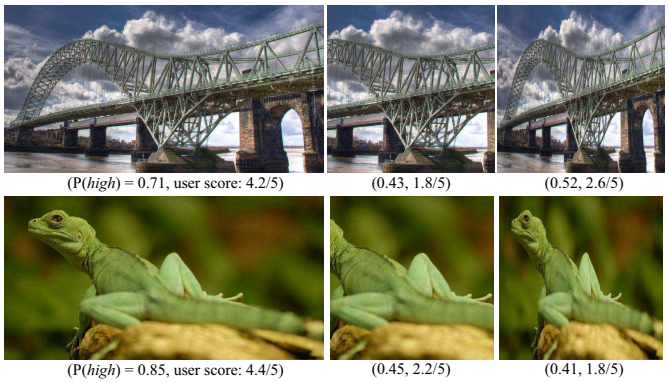
Table 2: Comparison between deep ConvNets with and without spatial pooling layers.

the gound-truth binary aesthetics labels. The ROC curve is obtained by varying the binarization threshold value from 0 to 1 and computing the true positive rate and false positive rate at each threshold. The area under the ROC curve (AUC) is computed to assess the performance of the network over a wide range of binarization threshold values. All the networks are trained using the training dataset obtained with $\delta = 0$. The results show that both SPP-CNN and MNA-CNN can significantly improve the aesthetics assessment performance over the three fixed-size networks. In addition, our multi-net based architecture MNA-CNN performs better than the spatial pyramid pooling based architecture SPP-CNN. As discussed earlier in Section 3.1, while the SPP-CNN architecture allows multi-scale processing to be performed by the upper fully connected layer of the network, it requires the learning process to capture the complex interaction among different lower pooling layers, which demands a larger amount of data and longer training time to learn successfully. Our MNA-CNN architecture, on the other hand, trains a sub-network for each scale and then aggregate them together, enabling easy training.

### 4.3. Composition-preserving Analysis

It is interesting to examine if our MNA-CNN network has learned to respond to the change in image composition, especially those caused by cropping and scaling. To test this, we collect 20 high-quality images from the AVA benchmark. For each original image, we generate a cropped version where the original image is cropped to its center using a square cropping windows whose side equals to the smaller dimension of the image, and a scaled version where the image is scaled along the longer side to make it square, as illustrated in Figure 6. In this way, we have 60 images in total. We then ask five users to rate each of these 60 images with an aesthetics score ranging from 1 to 5, with 5 indicating the highest aesthetics value. We randomize the order of these 60 images and show one image to each user at each time. We average the scores from the five users as the final score for each image. Finally, for each of the 20 original images, we pair it with one of the transformed images and obtain 40 pairs in total. We label a pair of images as *desc* if the transformed image is rated with a higher score than the original one and *asc* otherwise.

We then use our MNA-CNN method to rate each image $I$ with the output probability $P(I = high)$ and obtain the *desc* or *asc* label for the above 40 pairs according to the predicted scores. We found that for all these 40 pairs of original/transformed images, the labels from our method agree with the ground-truth labels computed from the user scores. This shows that our MNA-CNN is able to reliably respond to the change of image composition caused by cropping or scaling. On the other hand, the baseline deep networks VGG-crop, VGG-pad, and VGG-scale only agree with the user ratings for 59.5%, 61.5%, and 63.1%, respectively. Figure 6 shows two example images used in the study and their transformed versions, along with the average user given scores and our MNA-CNN predicted posteriors.

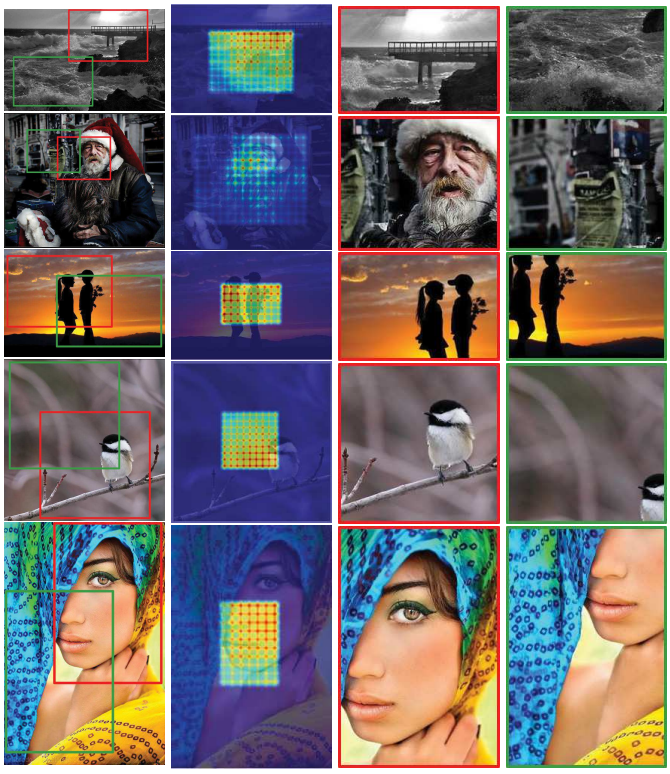



(a) Input image (b) Cropped version (c) Scaled version

Figure 6: Test on composition changes. Each input image is cropped and scaled to change the photo composition. The posterior predicted by our MNA-CNN method and the user score for each image is shown below each image. This test shows that our MNA-CNN method can reliably rate images with different compositions.

**Automatic cropping.** We conduct another study to visually evaluate how our method can be used to guide image cropping. Specifically, given an image, we slide a cropping window through the whole image with the step size of 20 pixels. We then employ our MNA-CNN method to predict the score for each cropping result. Figure 7 shows some example images and their highest- and lowest-scored cropping results in (c) and (d), respectively. We also create a quality map (b) by assigning the score of each cropping window to the image pixel corresponding to its center. The map is then smoothed for better visualization. The high values in the map indicate the locations on the image that our method suggests a cropping window should be centered at to create a high-quality cropping result.

It is interesting to note that our method not only tends to capture important content in the photo (e.g. the human face, the bird's head, and the bridge) but also learns where to position the main subject to create a well-composed cropping result. For example, even with a small cropping window, our method tends to select the cropping window such that the main subject is positioned a little off-center in the resulted image, which agrees with common photography techniques such as rule of thirds [18].



(a) Input Image (b) Quality Map (c) Highest quality (d) Lowest quality

Figure 7: Automatic cropping. We slide a cropping window through the whole image with the step size of 20 pixels. Each cropping result is scored by our MNA-CNN method. We show the highest rated cropping results in (c) and the lowest-rated cropping results in (d). (b) is a cropping quality map with high values indicating the locations on the image that our method suggests a cropping window should be centered at to create a good cropping result.

### 4.4. Scene-Aware MNA-CNN Performance

To examine the effectiveness of the scene-aware aggregation component, we train the MNA-CNN network on half of the available training data, and use the other half to train the scene-based aggregation model as described in Section 3.1.2. Table 1 shows that incorporating the scene category prediction information (MNA-CNN-Scene) improves the aggregation performance of our MNA-CNN network.

Figure 8 further explains the improvement of scene-based aggregation upon our MNA-CNN method. For each of the predicted seven scene categories, this figure shows the performance of each individual sub-network of the MNA-CNN architecture taken independently as well as the performance after aggregation. We can see that the relative performance among the sub-networks vary across different scene categories. Taking the scene prediction as augmented information can therefore help the aggregation model effectively combine the prediction from individual sub-networks to produce the better overall prediction.

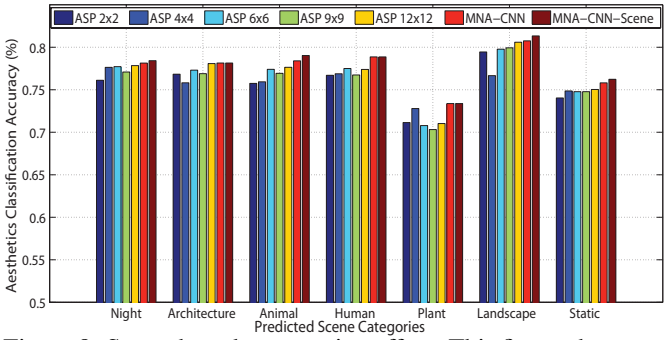



Figure 8: Scene-based aggregation effect. This figure shows the performance of each sub-network, the average aggregation result (MNA-CNN) and the scene-aware aggregation result (MNA-CNN-Scene).

**Discussion.** As expected, padding an image performs better than cropping, as reported in Table 2. However, it is interesting why padding performs slightly worse than scaling given that it does not crop off or distort image content. We suspect that there are two reasons. First, padding adds artificial boundaries between the real image and the padding areas and the boundaries vary over different images. This could confuse a deep learning algorithm. Second, padding is typically coupled with uniform scaling, which can make the small yet interesting content difficult to appreciate as shown in Figure 1 (c). While this is not directly related to our method, it will be interesting to study in the future.

Our MNA-CNN-Scene model is trained on all the training data and uses the scene information only at the top aggregation layer. Previous studies [24, 48] have found it beneficial to learn different aesthetics models specifically for each scene category. Training scene-specific deep neural networks is challenging as it requires a large amount of training data and very accurate scene categorization results. We plan to study this problem in our future work.

## 5. Conclusion

This paper presents a scene-aware Multi-Net Adaptive Spatial Pooling ConvNet (MNA-CNN) for photo aesthetics assessment. This MNA-CNN deep ConvNet is trained and tested with images at their original sizes and aspect ratios without first transforming them into a fixed size and thus preserves the aesthetics of the original images. This scene-aware MNA-CNN has three enabling features. First, it uses an adaptive spatial pooling layer upon regular convolutional and pooling layers. This adaptive spatial pooling layer has a fixed-size output while having a variable receptive field size to handle images with different sizes and aspect ratios. Second, it uses multiple sub-networks to capture aesthetics features at multiple scales. Finally, it uses a scene-aware aggregation layer to combine these sub-networks into a powerful one. Our experiments on the large-scale AVA benchmark show that our scene-aware MNA-CNN can significantly improve the state of the art in photo aesthetics assessment.

**Acknowledgment.** This work was supported by NSF IIS-1321119 and CNS-1218589.

## References


[1] G. Bertasius, J. Shi, and L. Torresani. Deepedge: A multi-scale bifurcated deep network for top-down contour detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 4

[2] S. Bhattacharya, R. Sukthankar, and M. Shah. A framework for photo-quality assessment and enhancement based on visual aesthetics. In *ACM International Conference on Multimedia*, pages 271–280, 2010. 1, 2

[3] T. Brandão and M. Queluz. No-reference quality assessment of h. 264/avc encoded video. *IEEE Transactions on Circuits and Systems for Video Technology*, 20(11):1437–1447, 2010. 2

[4] N. Damera-Venkata, T. Kite, W. Geisler, B. Evans, and A. Bovik. Image quality assessment based on a degradation model. *IEEE Transactions on Image Processing*, 9(4):636–650, 2000. 2

[5] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *European Conference on Computer Vision*, pages 288–301, 2006. 1, 2

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. 3

[7] S. Dhar, V. Ordonez, and T. Berg. High level describable attributes for predicting aesthetics and interestingness. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1657–1664, June 2011. 1, 2

[8] T. G. Dietterich. Ensemble methods in machine learning. In *Proceedings of the First International Workshop on Multiple Classifier Systems*, pages 1–15, 2000. 4

[9] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1915–1929, 2013. 4

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 2

[11] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(9):1904–1916, 2015. 2, 3, 4, 6

[12] W. Jiang, A. Loui, and C. Cerosaletti. Automatic aesthetic value assessment in photographic images. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 920–925, July 2010. 1, 2

[13] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Wang, J. Li, and J. Luo. Aesthetics and emotions in images. *Signal Processing Magazine, IEEE*, 28(5):94–115, Sept 2011. 1, 2

[14] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, and H. Winnemoeller. Recognizing image style. In *British Machine Vision Conference*, 2014. 2

[15] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015. 2

[16] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 2

[17] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 419–426, June 2006. 1, 2

[18] B. P. Krages. *The Art of Composition*. Allworth Communications, Inc., 2005. 7

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1106–1114, 2012. 2

[20] H. Kwon, Y.-W. Tai, and S. Lin. Data-driven depth map refinement via multi-scale sparse representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 4

[21] G. Li and Y. Yu. Visual saliency based on multiscale deep features. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 4

[22] X. Li. Blind image quality assessment. In *Proceedings of International Conference on Image Processing*, pages 449–453, 2002. 2

[23] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 4

[24] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Wang. Rating image aesthetics using deep learning. *IEEE Transactions on Multimedia*, 17(11):2021–2034, Nov 2015. 5, 6, 8

[25] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang. Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the ACM International Conference on Multimedia*, pages 457–466, 2014. 1, 2, 5, 6

[26] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Wang. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In *IEEE International Conference on Computer Vision*, 2015. 2, 5, 6

[27] Y. Luo and X. Tang. Photo and video quality evaluation: Focusing on the subject. In *European Conference on Computer Vision*, pages 386–399, 2008. 1, 2

[28] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. In *IEEE International Conference on Computer Vision*, pages 1784–1791, Nov 2011. 1, 2

[29] N. Murray, L. Marchesotti, and F. Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415, June 2012. 1, 5, 6

[30] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba. Considering temporal variations of spatial visual distortions in video quality assessment. *IEEE Journal of Selected Topics in Signal Processing*, 3(2):253–265, 2009. 2

[31] M. Nishiyama, T. Okabe, I. Sato, and Y. Sato. Aesthetic quality classification of photographs based on color harmony. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 33–40, June 2011. 1, 2

[32] Y. Niu and F. Liu. What makes a professional video? a computational aesthetics approach. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(7):1037–1049, July 2012. 1

[33] Y. Ou, Z. Ma, T. Liu, and Y. Wang. Perceptual quality assessment of video considering both frame rate and quantization artifacts. *IEEE Transactions on Circuits and Systems for Video Technology*, (99):1–1, 2010. 2

[34] W. Ouyang, X. Wang, X. Zeng, S. Qiu, P. Luo, Y. Tian, H. Li, S. Yang, Z. Wang, C.-C. Loy, and X. Tang. Deepid-net: Deformable deep convolutional neural networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2015. 2

[35] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2007. 1, 2

[36] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *European Conference on Computer Vision*, pages 143–156, 2010. 1, 2

[37] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015. Published online 2014; based on TR arXiv:1404.7828 [cs.NE]. 2

[38] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *International Conference on Learning Representations*. CBLS, April 2014. 2

[39] K. Seshadrinathan and A. Bovik. Motion tuned spatio-temporal quality assessment of natural videos. *IEEE Transactions on Image Processing*, 19(2):335–350, 2010. 2

[40] H. Sheikh, A. Bovik, and L. Cormack. No-reference quality assessment using natural scene statistics: Jpeg2000. *IEEE Transactions on Image Processing*, 14(11):1918–1927, 2005. 2

[41] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 2, 5

[42] 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 *ACM International Conference on Multimedia*, pages 1213–1216, 2011. 1, 2

[43] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 2

[44] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2

[45] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2015. 2

[46] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. June 2014. 2

[47] H. Tang, N. Joshi, and A. Kapoor. Blind image quality assessment using semi-supervised rectifier networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2877–2884, 2014. 2

[48] X. Tang, W. Luo, and X. Wang. Content-based photo quality assessment. *Multimedia, IEEE Transactions on*, 15(8):1930–1943, Dec 2013. 1, 2, 4, 5, 8

[49] H. Tong, M. Li, H. Zhang, C. Zhang, J. He, and W. Ma. Learning no-reference quality metric by examples. In *Proceedings of International Conference on Multimedia Modelling*, 2005. 2

[50] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 2

[51] O. Wu, W. Hu, and J. Gao. Learning to predict the perceived visual quality of photos. In *IEEE International Conference on Conputer Vision*, pages 225–232, Nov 2011. 1, 2

[52] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1912–1920, 2015. 2

[53] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, pages 3320–3328. 2014. 2

[54] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, pages 487–495. 2014. 2, 4, 5

[55] Z.-H. Zhou. *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall/CRC, 1st edition, 2012. 4