

000
001
002054
055
056

A-Lamp: Adaptive Layout-Aware Multi-Patch Deep Convolutional Neural Network for Photo Aesthetic Assessment

057
058
059003
004
005
006
007
008
009
010
011
012
013
014
015060
061
062
063
064
065
066
067

Anonymous CVPR submission

068
069
070
071
072
073
074

Paper ID 1857

075
076
077
078
079

Abstract

080

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 image's 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 fine grained details and holistic image layout simultaneously. To enable training on these hybrid inputs, we extend the method by developing a dedicated double-subnet neural network structure, i.e. a Multi-Patch subnet and a Layout-Aware subnet. We further construct an aggregation layer to effectively combine the hybrid features from these two subnets. Extensive experiments on the large-scale aesthetics assessment benchmark (AVA) demonstrate significant performance improvement over the state-of-the-art in photo aesthetic assessment.

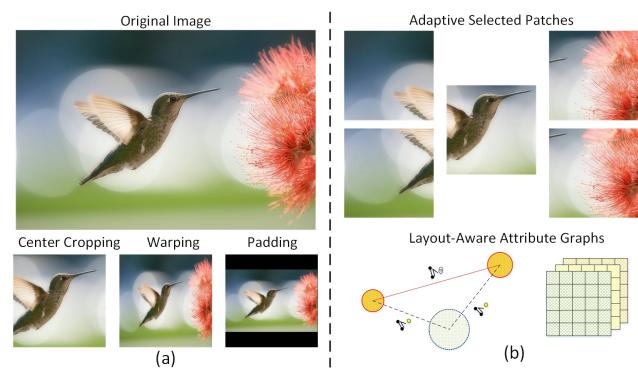
081
082
083
084

Figure 1. Conventional CNN methods (a) transform images via cropping, warping and padding. The proposed A-Lamp CNN (b) takes multiple patches and attributes graphs as inputs to represent fine grained details and the overall layout.

085
086
087
088
089

more generic features[29, 34, 38] to predict photo aesthetics. However, these generic features may not be suitable for assessing photo aesthetics, as they are designed to capture the general characteristics of the natural images instead of describing the aesthetics of the images.

090
091
092
093
094

Because of the limitations of these feature-based approaches, many researchers have recently turned to use deep learning strategy to extract effective aesthetics features [23, 21, 28, 39, 14]. These deep CNN methods have indeed shown promising results. However, the performance is often compromised by the constraint that the neural network only takes the fixed-size input. To accommodate this requirement, input images will need to be obtained via cropping, warping, or padding. As we can see from Figure 1, these additional operations often alter image composition, reduce image resolution, or cause extra image distortion, and thus impair the aesthetics of the original images because of potential loss of fine grained details and holistic image layout. However, such fine-grained details and overall image layout are critical for the task of image quality assessment. He[9] and Mai[28] tried to address the limitation in fixed-size input by training images in a few different scales to mimic varied input sizes. However, they still learn

095
096
097

1. Introduction

098
099
100

Automatic image aesthetics assessment is challenging. Among the early efforts [5, 15], various hand-craft aesthetics features [25, 2, 40, 6, 38, 4] have been manually designed to approximate the photographic and psychological aesthetics rules. However, to design effective aesthetics features manually is still a challenging task because even the very experienced photographers use very abstract terms to describe high quality photos. Other approaches leveraged

101
102
103
104105
106
107

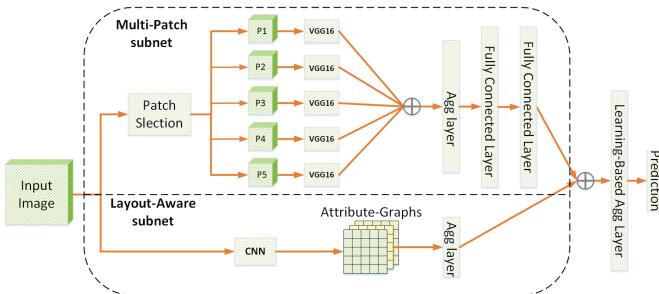


Figure 2. The architecture of the A-Lamp CNN. More detailed illustrations for Multi-Patch subnet and Layout-Aware subnet can be seen in Figure 3 and Figure 4.

from transformed images, which may result in substantial loss of fine grained details and undesired distortion of image layout.

Driven by this important issue, a question arises: *Can we simultaneously learn fine-grained details and the overall layout to address the problems caused by the fixed-size limitation?* To resolve this technical issues, we present in this paper a dedicated CNN architecture named A-Lamp. This novel scheme can accept arbitrary images with its native size. Training and testing can be effectively performed by considering both fine-grained details and image layout, thus preserving the information from original images.

Learning both fine-grained details and image layout is indeed very challenging. First, the detail information is contained in the original, high resolution images. Training deep networks with large-sized input dimensions requires much longer training time, training dataset, and hardware memory. To enable learning from fine grained details, a multi-patch-based method was proposed in [23]. This scheme shows some promising results. However, these randomly picked bag of patches cannot represent the overall image layout. In addition, this random cropping strategy requires a large number of training epochs to cover the desired diversity in training, which lead to low efficiency in learning.

Second, how to effectively describe specific image layout and incorporate it into the deep CNN is again very challenging. Existing works related to image layout descriptors are dominantly based on a few simple photography composition principles, such as visual balance, rule of thirds, golden ratio, and so on. However, these general photography principles are inadequate to represent local and global image layout variations. To incorporate global layout into CNN, transformed images via warping and center-cropping have been used to represent the global view [22]. However, such transformation often alters the original image composition or causes undesired layout distortion.

In this paper, we resolved this challenges by developing an Adaptive Layout-Aware Multi-Patch Convolutional Neural Network (A-Lamp CNN) architecture. The design A-Lamp is inspired jointly by the success of fine-grained detail

learning using multi-patch strategy [23, 20] and the success of holistic layout representation by attribute graph. However, the proposed scheme can successfully overcome the stringent limitations of the existing schemes. Like DMA-Net in [23], our proposed A-Lamp CNN also crops multiple patches from original images to preserve fine-grained details. Comparing to DMA-Net, this scheme has two major innovations. First, instead of cropping patches randomly, we propose an adaptive multi-patch selection strategy. The central idea of adaptive multi-patch selection is to maximize the input information more efficiently. We achieve this goal by dedicatedly selecting the patches that play important role in affecting images' aesthetics. We expect that the proposed strategy shall be able to outperform the random cropping scheme even with substantially less training epochs. Second, unlike the DMA-Net that just focus on the fine-grained details, this A-Lamp CNN incorporates the holistic layout via the construction of attribute graph. We use graph nodes to represent objects and the global scene in the image. Each object (note) is described using object-specific local attributes while the overall scene is represented with global attributes. The combination of both local and global attributes captures the layout of an image effectively. This attribute-graphs based approach is expected to model image layout more accurately and outperform the existing approaches based on warping and center cropping. These two innovations result improvement in both efficiency and accuracy over DMA-Net. The main contributions of this proposed A-Lamp scheme can be summarized into three-fold:

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

2. Related Work

2.1. Deep Convolutional Neural Networks

Deep learning methods have shown great successes in various computer vision tasks, including conventional tasks in object recognition [43], object detection [9, 20], and image classification [36, 10], as well as contemporary tasks in image captioning [1], saliency detection [32], style recog-

162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215

216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
nition [8, 14] and photo aesthetics assessment [21, 23, 39, 28, 13]. Most existing deep learning methods transform input images via cropping, scaling, and padding to accommodate the deep neural network architecture requirement in fixed size input which would compromise the network performance as we have discussed previously.

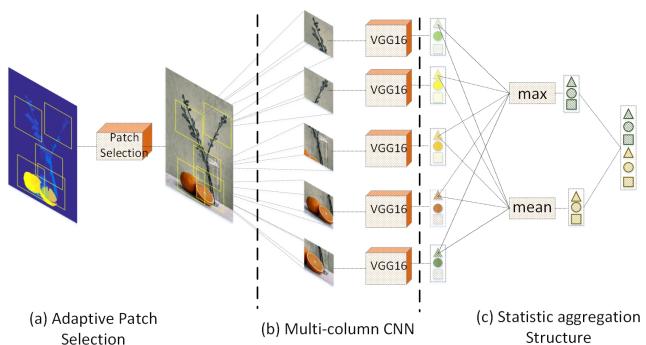
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
Recently, new strategies to construct adaptive spatial pooling layers have been proposed to alleviate the fixed-size restriction [9, 28]. In theory, these network structures can be trained with standard back-propagation, regardless of the input image size. In practice, the GPU implementations of deep learning are preferably run on fixed input size. The recent research [9, 28] mimic the variable input sizes by using multiple fixed-size inputs which are obtained by scaling from original images. It is apparently still far from arbitrary size input. Moreover, the learning is still from transformed images, which inherently compromise the performance of the deep learning networks.

250
251
252
253
254
255
256
257
258
259
260
Others have proposed dedicated network architectures. A double-column deep convolutional neural network was developed in [21] to support heterogeneous inputs with both global and local views. The global view is represented by padded or warped image while, the local view is represented by randomly cropped single patch. This work was further improved in [23], where a deep multi-patch aggregation network was developed (DMA-Net) to take multiple randomly cropped patches as input. This network have shown some promising results. However, these random order of bag of patches is unable to capture image layout information, which is crucial in image aesthetics assessment. Furthermore, to ensure that most of the information will be captured by the network, this scheme uses large number of randomly selected groups of patches for each image, and trains them for 50 epochs, resulting in very low training efficiency.

2.2. Image Layout Representation

261
262
263
264
265
266
267
268
269
To represent holistic image layout, existing works [19, 31, 33, 41, 45, 26, 27, 48] adopt dominantly the model of image composition by approximating some simple traditional photography composition guidelines, such as visual balance, rule of thirds, golden ratio, and diagonal dominance. However, these heuristic guidance-based descriptors cannot capture the intrinsic of photo aesthetics in terms of image layout.

270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
Attribute-graph, which has long been used by the vision community to represent structured groups of objects [7, 24, 12, 37, 47], shows promising results in representing complicated image layout. The spatial relationship between a pair of objects was considered in [18] even though the overall geometrical layout of all the objects and the object characteristics cannot be accounted for with this method. The scheme reported in [42] was able to maintain spatial relationships among objects but related background informa-



295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
Figure 3. The architecture of Multi-Patch subnet: (a) adaptive patch selection module, (b) a set of paralleled shared weights CNNs that are used for extracting deep features from each of the patch, (c) aggregation structure which combines the extracted deep features from the multi-column CNNs jointly.

tion and object attributes were not addressed. The scheme reported in [17] considers both objects and their interrelations, but have not been integrated with the holistic background modeling. The scheme in [3] performs image aesthetics ranking by constructing the triangular object structures with attribute features. However, this scheme lacks of proper account for the global scene context.

3. Adaptive Layout-Aware Multi-Patch CNN

295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
The architecture of the proposed A-Lamp is shown in Figure 2. Given an arbitrary sized image, multiple patches will be adaptively selected by the *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 *Attribute-Graphs*. At the end, a 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.

3.1. Multi-Patch subnet

312
313
314
315
316
317
318
319
320
321
We represent each image with a set of carefully cropped patches, and associate the set with the image's label. The training data is $\{P_n, y_n\}_{n \in [1, N]}$, where $P_n = \{p_{nm}\}_{m \in [1, M]}$ is the set of M patches cropped from each image. The architecture of proposed Multi-Patch subnet is shown in Figure 3 and more details will be explained in this section.

3.1.1 Adaptive Patch Selection

322
323
Different from the random-cropping method in [23], we aim to carefully select the most discriminative and informative

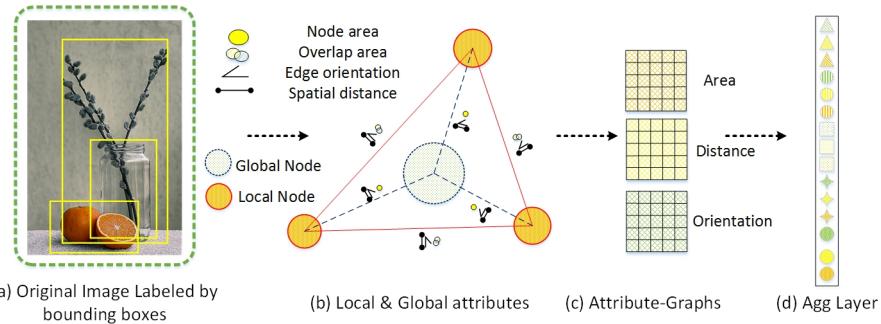


Figure 4. Pipeline of attribute-graphs construction. (a) Salient objects (labeled by yellow bounding boxes) are first detected by a trained CNN, and regarded as local nodes. The dashed green bounding box denote the overall scene, which served as a global node. (b) Local and global attributes are extracted from these nodes to capture the object topology and the image layout. (c) Attribute-graphs are constructed and (d) concatenated into an aggregation layer.

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:

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.

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 heavily depends 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.

Overlapping Constraint. Spatial distance among any patch pairs should also be considered to constrain 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:

$$\{c^*\} = \underset{i,j \in [1,M]}{\operatorname{argmax}} F(S, D_p, D_s) \quad (1)$$

$$F(\cdot) = \sum_{i=1}^M S_i + \sum_{i \neq j}^M D_p(\tilde{N}_i, \tilde{N}_j) + \sum_{i \neq j}^M D_s(c_i, c_j) \quad (2)$$

where $\{c_m^*\}_{m \in [1,M]}$ is the centers of the optimal set of M selected patches. $S_i = \frac{\text{sal}(p_i)}{\text{area}(p_i)}$ is the normalized saliency value for each patch p_i . The saliency value is obtained by a graph-based saliency detection approach [44]. $D_p(\cdot)$ is the pattern distance function which measures the difference of two patches’ patterns. Here we adopt edge and chrominance distribution to represent the pattern of each patch. Specifically, we model the pattern of a patch p_m using a multi-variate Gaussian:

$$\tilde{N}_m = \{\{N_e(\mu_e, \Sigma_e)\}_m, \{N_c(\mu_c, \Sigma_c)\}_m\}_{m \in [1,M]} \quad (3)$$

where $\{N_e(\mu_e, \Sigma_e)\}_m$ and $\{N_c(\mu_c, \Sigma_c)\}_m$ denote edge distribution and chrominance distribution of patch p_m , respectively. Σ_e and Σ_c are the covariance matrices of N_e and N_c . Therefore, measuring pattern difference between a pair of patches can be formulated by mapping these distributions \tilde{N}_m to the *Wasserstein Metric space* $M_{m \times m}$, and calculate the 1_{st} *Wasserstein distance* between \tilde{N}_i and \tilde{N}_j on this given metric space M . Following F. Pitie [35], the closed form solution is given by:

$$D_p(\cdot) = \Sigma_i^{-1/2} \left(\Sigma_i^{1/2} \Sigma_j \Sigma_i^{1/2} \right) \Sigma_i^{-1/2} \quad (4)$$

$D_s(\cdot)$ is the spatial distance function, which is measured by Euclidean Distance.

3.1.2 Orderless Aggregation Structure

We also perform the aggregation of the multiple instances to enable the proposed network learn from multiple patches cropped from a given image. Let $\langle \text{Blob}_n \rangle_l = \{b_i^n\}_{i \in [1,M]}^l$ be the set of patch features extracted from n_{th} image at l_{th} layer of the shared CNNs. $b_{i,l}^n$ is a K dimensional vector. T_k denotes the set of values of the k_{th} component of all $b_{i,l}^n \in \langle \text{Blob}_n \rangle_l$. For simplicity, we omit image index n and layer index l , thus $T_k = \{d_{ik}\}_{i \in [1,M]}$. The aggregation layer is comprised of a collection of statistical functions, i.e., $F_{Agg}^u = \{F_{Agg}^u\}_{u \in [1,U]}$. Each F_{Agg}^u

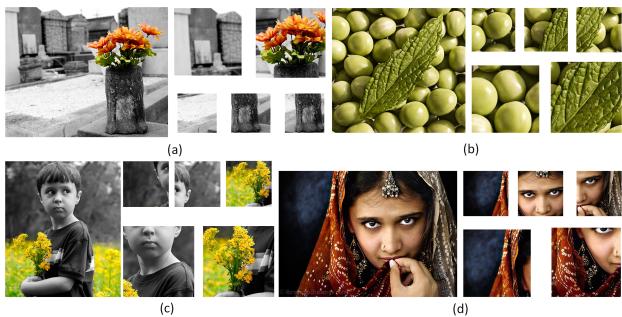


Figure 5. Examples of selected patches by the proposed Adaptive Patch-Selection scheme. In each group, original image is on the left side, and patches are located on the right side. We zoom in the patches that have more details for clear display. In practice, the size of all the patches are 224×224 .

computes *Blob* returned by the shared CNNs. Here we adopt a modified statistical functions proposed in [23], i.e. $U = \{\max, \text{mean}\}$ ¹. The outputs of the functions in U are concatenated to produce a K_{stat} -dimensional feature vectors. Two fully connected layers are followed to implement multi-patch aggregation component. The whole structure can be expressed as a function $f : \{\text{Blob}\} \rightarrow K_{\text{stat}}$:

$$f(\text{Blob}) = W \times (\oplus_{u=1}^U \oplus_{k=1}^K F_{\text{Agg}}^u(T_k)) \quad (5)$$

where \oplus is a vector concatenation operator which produces a column vector, $W \in K_{\text{stat}} \times UK$ is the parameters of the fully-connected layer. Figure 3 shows an example of Statistics Aggregation Structure with $M = 5$ and $K = 3$. In practice, the feature dimension $K = 4096$.

3.2. Layout-Aware Subnet

We first employ a trained CNN [46] to localize the salient objects. Let $I : \{B_i, s_i\}_{N_{\text{obj}}}$ 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 s_i , N_{obj} 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. Each object presents in the image contributes to a graph node resulting in a total of N_{obj} local nodes $V_l = \{v_1, \dots, v_{N_{\text{obj}}}\}$. local edges E_l refer to the edges between a pair of local nodes, there will be $(N_{\text{obj}} - 1)!$ 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 attributes capture the relative arrangement of the ob-

¹Through extensive experiments, we find that max, min showing the best performance. The statistical functions adopted in [23], i.e. min, max, mean, median, not result in performance improvement, and even worse because the potential of over-fitting caused by the too large vector dimension.

jects with respect to each other, which are represented by

$$\Phi_l(i, j) = \{dist(i, j), \theta(i, j), \hat{o}(i, j)\}_{v_i, v_j \in V_l} \quad (6)$$

where $\Phi_l(i, j)$ represents the attribute of a pair of connecting node v_i and v_j . $dist(i, j)$ is the spatial distance between object centroids. $\theta(i, j)$ represents the angle of the graph edge with respect to the horizontal taken in the anti-clockwise direction. It indicates the relative spatial organization of the two objects. $\hat{o}(i, j)$ represents the amount of overlap between the bounding boxes of the two objects and is given by

$$\hat{o}_{ij} = \frac{area(v_i) \cap area(v_j)}{\min(area(v_i), area(v_j))} \quad (7)$$

where $area(v_i)$ is the fraction of the image area occupied by the i^{th} 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. The global node V_g represents the overall scene. The edges connecting local nodes and global node are global edges E_g , there will be N_{obj} such edges. The global node captures the overall essence of the image. The global attributes Φ_g are given by

$$\Phi_g(i, g) = \{dist(i, g), \theta(i, g), area(v_i)\}_{v_i \in V_l, v_g \in V_g} \quad (8)$$

where $dist(i, g)$ and $\theta(i, g)$ are the magnitude and orientation of the edge connecting the centroid of the object corresponding to node v_i to the global centroid c_g . The edges connecting each object to the global node illustrate the placement of that object with respect to the overall object topology.

An aggregation layer is adopted to concatenate the constructed attribute graphs into a feature vector $\vec{\nu}$, and further combined with the Multi-Patch subnet, which is illustrated in Figure 2.²

4. Experimental Results

In the implementation, we release the memory burden 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. VGG16 is one of the state-of-the-art object-recognition networks that is pre-trained on the ImageNet [16]. Following Lu [23], The number of patches in a bag is set to be 5. The patch size is fixed to be $224 \times 224 \times 3$. The base learning rate is 0.01, the weight decay is 1e-5 and momentum is 0.9.

²By statistical study, we find that, the confidence score is very low when $N_{\text{obj}} \geq 5$. So we set $N_{\text{obj}} = 4$ to fix the feature vector $\vec{\nu}$ dimension.

Method	Accuracy
DMA-Net _{ave}	73.1 %
DMA-Net _{max}	73.9 %
DMA-Net _{stat}	75.4%
DMA-Net _{fc}	75.4%
Random-MP-Net	74.8%
New-MP-Net	81.7%

Table 1. Performance comparisons of Adaptive Multi-Patch subnet with other multi-patch-based CNNs.

Method	Accuracy	F-measure
AVA	67.0 %	<i>NA</i> *
VGG-Center-Crop	72.2 %	0.83
VGG-Wrap	74.1 %	0.84
VGG-Pad	72.9 %	0.83
SPP-CNN	76.0 %	0.84
MNA-CNN	77.1 %	0.85
MNA-CNN-Scene	77.4 %	<i>NA</i> *
DCNN	73.25 %	<i>NA</i> *
DMA-Net-ImgFu	75.4 %	<i>NA</i> *
New-MP-Net	81.7%	0.91
A-Lamp	82.5 %	0.92

Table 2. Comparisons of A-lamp with the state-of-the-art. * These results are not reported in the original papers [23, 22, 28, 30].

All the network training and testing are done by using the Caffe deep learning framework[11].

We systematically evaluate the proposed scheme on the AVA dataset [30], 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 [21, 23, 28, 30] in which roughly 20,000 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 or equal to 5 are labeled as low quality and those with mean ratings larger than 5 are labeled as high quality.

4.1. 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 it with some other multi-patch-based networks.

DMA-Net. DMA-Net proposed in [23] is a very recent dedicated deep Multi-Patch-based 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. Here we compare New-MP-Net with four types of DMA-Net architecture. DMA-Net_{ave} and DMA-Net_{max} train the DMA-Net using standard patch pooling scheme, where DMA-Net_{ave} performs average pooling and DMA-Net_{max} performs max pooling. The DMA-Net using Statistics Aggregation Structure is denoted as DMA-Net_{stat} and Fully-Connected Sorting Aggregation Structure as DMA-Net_{fc}.

MP-Net. The Multi-Patch subnet that takes the inputs by our proposed adaptive patch selection scheme is denoted as **New-MP-Net**. Since we adopt much deeper shared column CNNs (VGG16) in **New-MP-Net**. One may argue that the better performance may rely on the adoption of VGG16. Therefore, we train and test **New-MP-Net** by the same random cropping strategy in [23], which is denoted as Random-MP-Net. Specifically, we randomly crop 50 groups of patches from the original image with a 224×224 cropping window. For each testing image, we perform prediction for 50 random crops and take their average as the final prediction result.

The experimental results are shown in Table 1. We can see that, New-MP-Net outperforms all types of DMA-Net architectures. Although DMA-Net randomly cropped 50 groups of patches to train, and the total training has 50 epochs. The randomness in cropping was not able to effectively capture useful information and may cause the training to be confusing for the network. Besides, we find that most of the random generated patches are cropped from the same location of the image. That means, there are a large number of data repeatedly fed into the network, thus lead to the risk of over-fitting. Comparing the accuracy and F-measure of New-MP-Net (81.7% and 0.91) with Random-MP-Net (71.2% and ~~F-measure~~ 0.83), we can see that even using the same network architecture, ~~the performance is impaired by using random cropping strategy~~.

4.2. Effectiveness of Adaptive Patch Selection

Instead of random cropping, we adaptively select the most informative and discriminative patches as input, which is the key to achieve substantial performance enhancement. From Figure 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 which shows different patterns, i.e. the blue sky and the green ground, have also been selected. Therefore, the global characteristics, e.g. color harmony, Low-of-Depth, can also be perceived by learning these patches jointly. More examples of selected patches are shown in Figure 5. 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, face and the orange flowers), but also is capa-

648 ble of capturing the pattern diversity (e.g. the green leaf and
649 green beans, the flower and the gray wall). Furthermore, the
650 proposed adaptive patch selection strategy is also able to en-
651 hance the training efficiency. The result of New-MP-Net is
652 obtained by taking 20-30 training epochs, substantially less
653 than 50 epochs reported in [23], while still achieving better
654 performance.
655

4.3. Comparison with the State-of-the-Art

656 Table 2 shows the results of the proposed A-Lamp CNN
657 on the AVA dataset [30] for image aesthetics categorization.
658 The AVA dataset provides the state-of-the-art results for
659 methods that use manually designed features and generic
660 image features for aesthetics assessment. It is obvious that,
661 all recently developed deep CNN schemes outperform these
662 conventional feature-based approaches.
663

664 **A-Lamp vs. Baseline.** To examine the effectiveness
665 of the proposed scheme, we compare New-MP-Net and A-
666 Lamp with some baseline methods that take only fixed-size
667 inputs. In particular, we experiment on VGG16 with three
668 types of transformed inputs. The input of VGG16-Center-
669 Crop is obtained by cropping from the center of the orig-
670 inal image with a 224×224 cropping window. The input
671 of VGG16-Wrap is obtained by scaling the original input
672 image to the fixed size of 224×224 . In the experiment of
673 VGG16-Pad, the original image is uniformly resized such
674 that the larger dimension becomes 224 and the aspect ratio
675 is preserved. The 224×224 input is then formed by
676 padding the remaining dimension of the transformed im-
677 age with zero-valued pixels. We can see from Table 2 that,
678 both New-MP-Net and A-Lamp outperform these fixed-size
679 input VGG nets. Such results confirmed that training net-
680 work on multiple patches produces better prediction than
681 networks training on a single patch.
682

683 **A-Lamp vs. Non-fixed-Size CNNs.** We also compared
684 the proposed scheme with some latest non-fixed size re-
685 striction schemes, i.e. SPP-CNN [9] and MNA-CNN [28].
686 Different from these schemes that their inputs are from several
687 different level of scaled images, we implement the A-Lamp
688 network to be trained from the original images. The results
689 confirm that learning from original images is essential for
690 aesthetic assessment, as we have discussed earlier. In addi-
691 tion, higher prediction accuracy of the proposed scheme fur-
692 ther proves that, the *adaptive Multi-Patch strategy* is more
693 efficient than the spatial pyramid pooling strategy adopted
694 in SPP-CNN and MNA-CNN.
695

A-Lamp vs. Layout-Aware CNNs .

696 To show the effectiveness of the proposed layout-
697 aware subnet, we compare A-Lamp with several latest
698 deep CNN networks that incorporate global information for
699 learning. MNA-CNN-Scene [28] replace the average op-
700 erator in the MNA-CNN network with a new aggregation
701 layer that takes the concatenation of the sub-network pre-



Figure 6. Prediction results on transformed images. Images from left to right are original ones, down sampled version and warped version. We zoom in some regions for comparison the details of original images and the down sampled images

dictions 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.

ii. DCNN [22] is a double column convolutional neural network which combines random cropped and warped images as inputs to perform training . 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 images to capture local and global image characters is not as effective as our approach.

iii. The result of DMA-Net-ImgFu (75.4 %) [23] is obtained by averaging the prediction results of DMA-Net and the fine tuned Alexnet [16]. It is interesting that, though they incorporated transformed entire images to represent global information, it still fall behind the performance of our proposed A-Lamp (82.5 %). Such results further validate the effectiveness of our proposed layout-aware subnet.

4.4. A-Lamp Effectiveness Analysis

From Table 2, we can see that, 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.

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 *randomly* collect 20 high quality images from the AVA dataset. We generate a down sampled version and a warped version from *the original image*. The down-sampled version keeps the same aspect ratio (i.e. the layout has not been 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

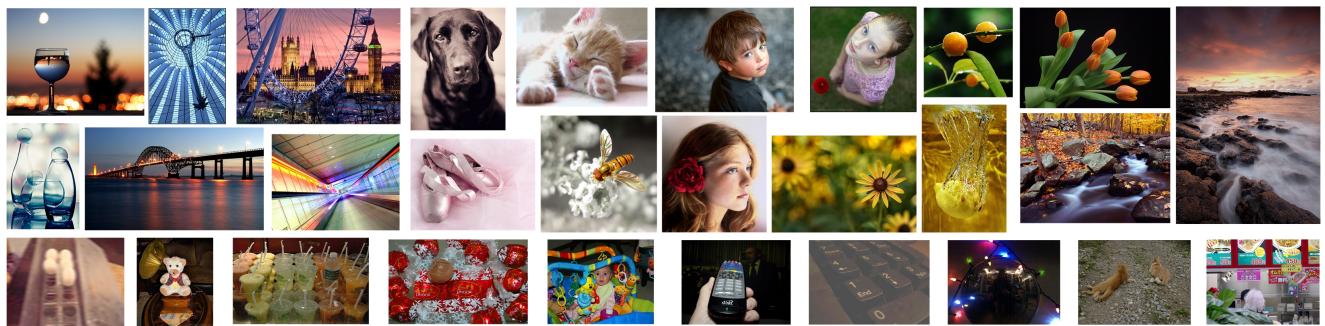


Figure 7. Results of predicted photos. The top two rows are predicted photos with high aesthetic scores. We random select these photos from eight categories [30]. The low aesthetic quality photos are shown in the third row.

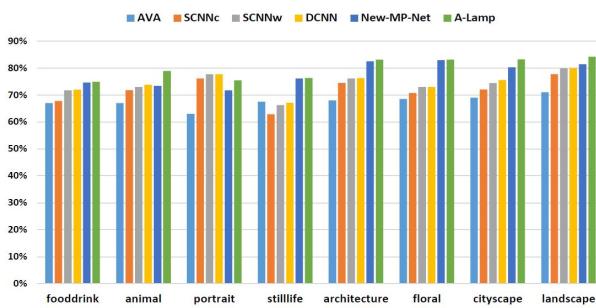


Figure 8. Comparison of aesthetic prediction performance in different content-based categories.

score for the original image than both transformed versions. Figure 6 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 of 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.

4.5. Content-based photo aesthetic analysis

To carry out content-based photo aesthetic study, we take photos in eight most popular semantic tags used in [30]: portrait, animal, still-life, food-drink, architecture, floral, cityscape and landscape. We used the same testing image collection used in [22], 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 [30] (denoted by

AVA) and the state-of-the-art approach in [22]. Specifically, SCNN_c and SCNN_w denote the single-column CNN in [22] that takes center cropping and warping, respectively, as inputs. DCNN denotes the double-column CNN in [22]. As shown in Figure 8, 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 Figure 7. 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. Figure 7 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.

5. Conclusion

This paper presents an Adaptive Layout-Aware Multi-Patch Convolutional Neural Network (A-Lamp) architecture for photo aesthetic assessment. This novel scheme is able to accept arbitrary sized images and to capture intrinsic aesthetic characteristics from both fine 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. Meanwhile, the proposed A-Lamp CNN can be directly applied to many other computer vision tasks, such as style classification, object recognition, image retrieval, and scene classification, which we leave as our future work.

864

865

866

867

868

869

870

871

872

873

874

875

876

877

878

879

880

881

882

883

884

885

886

887

888

889

890

891

892

893

894

895

896

897

898

899

900

901

902

903

904

905

906

907

908

909

910

911

912

913

914

915

916

917

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. ii
- [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. i
- [3] 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. iii
- [4] 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. i
- [5] 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. i
- [6] 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. i
- [7] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *Int. J. Comput. Vision*, 59(2):167–181, Sept. 2004. iii
- [8] 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. iii
- [9] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. volume abs/1406.4729, 2014. i, ii, iii, vii
- [10] 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. ii
- [11] 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. vi
- [12] 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. iii
- [13] 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. iii
- [14] S. Karayev, A. Hertzmann, H. Winnemoeller, A. Agarwala, and T. Darrell. Recognizing image style. *CoRR*, abs/1311.3715, 2013. i, iii
- [15] 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. i
- [16] 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. v, vii
- [17] 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. iii
- [18] 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. iii
- [19] L. Liu, R. Chen, L. Wolf, and D. Cohen-Or. Optimizing Photo Composition. *Computer Graphics Forum*, 2010. iii
- [20] 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. ii
- [21] 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. i, iii, vi
- [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. ii, vi, vii, viii
- [23] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang. Deep multi-patch 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. i, ii, iii, v, vi, vii
- [24] 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. iii
- [25] 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. i
- [26] S. Ma, Y. Fan, and C. W. Chen. Finding your spot: A photography suggestion system for placing human in the scene. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 556–560, Oct 2014. iii

918

919

920

921

922

923

924

925

926

927

928

929

930

931

932

933

934

935

936

937

938

939

940

941

942

943

944

945

946

947

948

949

950

951

952

953

954

955

956

957

958

959

960

961

962

963

964

965

966

967

968

969

970

971

- 972 [27] S. Ma, Y. Fan, and C. W. Chen. Pose maker: A pose recom- 1026
973 mendation system for person in the landscape photographing. In *Proceedings of the 22Nd ACM International Conference 1027
974 on Multimedia*, MM '14, pages 1053–1056, New York, 1028
975 NY, USA, 2014. ACM. **iii** 1029
- 976 [28] L. Mai, H. Jin, and F. Liu. Composition-preserving deep 1030
977 photo aesthetics assessment. In *The IEEE Conference on 1031
978 Computer Vision and Pattern Recognition (CVPR)*, June 1032
979 2016. **i, iii, vi, vii** 1033
- 980 [29] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka. 1034
981 Assessing the aesthetic quality of photographs using generic 1035
982 image descriptors. In *Proceedings of the 2011 International 1036
983 Conference on Computer Vision*, ICCV '11, pages 1784– 1037
984 1791, Washington, DC, USA, 2011. IEEE Computer Society. 1038
985 **i** 1039
- 986 [30] N. Murray, L. Marchesotti, and F. Perronnin. Ava: A large- 1040
987 scale database for aesthetic visual analysis. In *Computer 1041
988 Vision and Pattern Recognition (CVPR), 2012 IEEE Confer- 1042
989 ence on*, pages 2408–2415. IEEE, 2012. **vi, vii, viii** 1043
- 990 [31] P. Obrador, L. Schmidt-Hackenberg, and N. Oliver. The role 1044
991 of image composition in image aesthetics. In *2010 IEEE 1045
992 International Conference on Image Processing*, pages 3185– 1046
993 3188, Sept 2010. **iii** 1047
- 994 [32] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, and N. E. 1048
995 O'Connor. Shallow and deep convolutional networks for 1049
996 saliency prediction. In *The IEEE Conference on Computer 1050
997 Vision and Pattern Recognition (CVPR)*, June 2016. **ii** 1051
- 998 [33] J. Park, J. Y. Lee, Y. W. Tai, and I. S. Kweon. Modeling photo 1052
999 composition and its application to photo re-arrangement. In 1053
1000 *2012 19th IEEE International Conference on Image Process- 1054
1001 ing*, pages 2741–2744, Sept 2012. **iii** 1055
- 1002 [34] F. Perronnin, J. Sánchez, and T. Mensink. Improving the 1056
1003 fisher kernel for large-scale image classification. In *Pro- 1057
1004 ceedings of the 11th European Conference on Computer 1058
1005 Vision: Part IV*, ECCV'10, pages 143–156, Berlin, Heidelberg, 1059
1006 2010. Springer-Verlag. **i** 1060
- 1007 [35] F. Pitie and A. Kokaram. The linear monge-kantorovich linear 1061
1008 colour mapping for example-based colour transfer. In *Vi- 1062
1009 sual Media Production, 2007. IETCVMP. 4th European Con- 1063
1010 ference on*, pages 1–9, Nov 2007. **iv** 1064
- 1011 [36] S. Reed, Z. Akata, H. Lee, and B. Schiele. Learning deep 1065
1012 representations of fine-grained visual descriptions. In *The 1066
1013 IEEE Conference on Computer Vision and Pattern Recog- 1067
1014 nation (CVPR)*, June 2016. **ii** 1068
- 1015 [37] J. Shi and J. Malik. Normalized cuts and image segmenta- 1069
1016 tion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888– 1070
1017 905, Aug. 2000. **iii** 1071
- 1018 [38] H.-H. Su, T.-W. Chen, C.-C. Kao, W. H. Hsu, and S.- 1072
1019 Y. Chien. Scenic photo quality assessment with bag of 1073
1020 aesthetics-preserving features. In *Proceedings of the 19th 1074
1021 ACM International Conference on Multimedia*, MM '11, 1075
1022 pages 1213–1216, New York, NY, USA, 2011. ACM. **i** 1076
- 1023 [39] H. Tang, N. Joshi, and A. Kapoor. Blind image quality 1077
1024 assessment using semi-supervised rectifier networks. In *Pro- 1078
1025 ceedings of the 2014 IEEE Conference on Computer Vision 1079
1026 and Pattern Recognition*, CVPR '14, pages 2877–2884, 1027
1027 Washington, DC, USA, 2014. IEEE Computer Society. **i, iii**