A Deep Network Solution for Attention and Aesthetics Aware Photo Cropping

Wenguan Wang, Jianbing Shen, *Senior Member, IEEE*, and Haibin Ling

Abstract—We study the problem of photo cropping, which aims to find a cropping window of an input image to preserve as much as possible its important parts while being aesthetically pleasant. Seeking a deep learning-based solution, we design a neural network that has two branches for attention box prediction (ABP) and aesthetics assessment (AA), respectively. Given the input image, the ABP network predicts an attention bounding box as an initial minimum cropping window, around which a set of cropping candidates are generated with little loss of important information. Then, the AA network is employed to select the final cropping window with the best aesthetic quality among the candidates. The two sub-networks are designed to share the same full-image convolutional feature map, and thus are computationally efficient. By leveraging attention prediction and aesthetics assessment, the cropping model produces high-quality cropping results, even with the limited availability of training data for photo cropping. The experimental results on benchmark datasets clearly validate the effectiveness of the proposed approach. In addition, our approach runs at 5 *fps*, outperforming most previous solutions.

Index Terms—Photo cropping, attention box prediction, aesthetics assessment, deep learning.

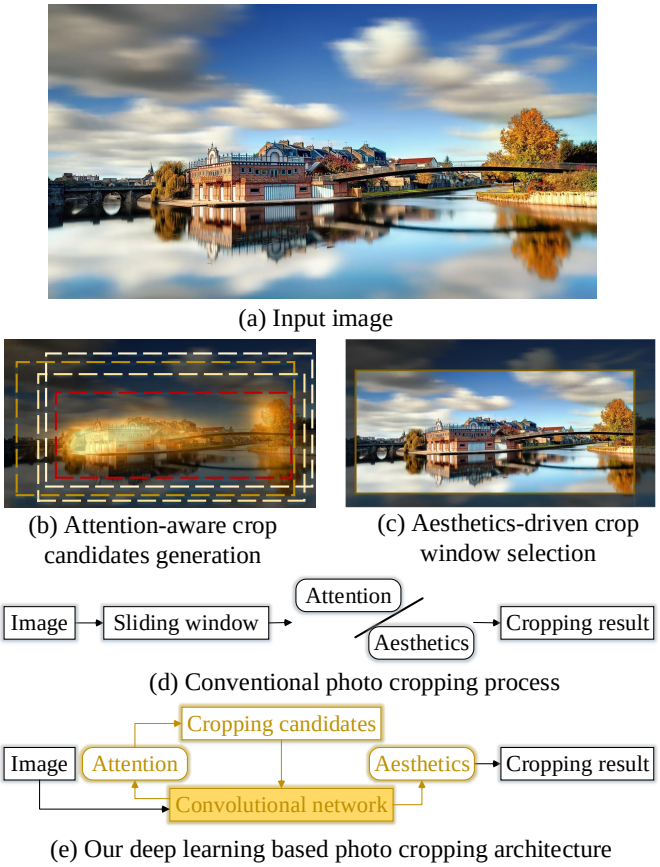
1 Introduction

1.1 Problem Statement and Motivation

Given an input photo, what is the best way to crop it? The answer, not surprisingly, varies from person to person, and even from time to time for the same person. In this paper, we study the problem in the general setting without prior knowledge of specific applications. In such setting, it is natural to expect a good cropping window to have two properties: keeping most of the important portion and being aesthetically pleasant. The idea can be viewed from the example in Figure 1.

The above general idea naturally inspires a photo cropping strategy through *determining-adjusting*. That is, one can first define a cropping window that covers the important region, and then adjust (iteratively) the position, size and ratio of the initial cropping until the satisfying result is achieved. This cropping strategy brings two advantages: (1) consideration of both image importance and aesthetics in a cascaded way; and (2) high computation efficiency since the searching space of the best cropping is limited to the neighborhood of the initial one.

Interestingly, however, most previous cropping approaches work differently. They usually generate a large number of sliding windows by varying sizes and aspect ratios over all the positions, and find the optimal cropping window by computing attention scores for all windows [1], [2], [3], or by analyzing their aesthetics [4], [5]. This *sliding-judging* strategy, as depicted in Figure 1 (d), is of high computation load, since its searching space spans all possible sub-windows of the entire photo. By contrast, the



(a) Input image

(b) Attention-aware crop candidates generation

(c) Aesthetics-driven crop window selection

(d) Conventional photo cropping process

(e) Our deep learning based photo cropping architecture

Fig. 1. (a) An input photo to be cropped. (b) The predicted attention box (red) and cropping candidates generated from it (yellow). (c) The final cropping with the maximum estimated aesthetic value. (d) Conventional image cropping methods with sliding-judging cropping strategy, which is time-consuming and violates natural cropping procedure. (e) Our algorithm as a cascade of attention-aware candidate generation and aesthetics-based cropping selection, which handles photo cropping more naturally via a unified neural network.

- *W. Wang and J. Shen are with Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology. (Email: wenguanwang.ai@gmail.com, shenjianbing@ucla.edu)*
- *H. Ling is with the Department of Computer and Information Sciences, Temple University, Philadelphia, PA, USA. (Email: hbling@temple.edu)*
- *Corresponding author: Jianbing Shen*

determining-adjusting strategy is more efficient by arranging the two key components sequentially and reduce the size of searching space.

Different than many previous approaches, in this paper, we design a deep learning-based photo cropping algorithm following the determining-adjusting strategy. Our algorithm models photo cropping as a cascade of attention bounding box *regression* and aesthetics *classification*. In particular, our model first determines an attention box that covers the most visually important area (the red rectangle in Figure 1 (b)), thus provides an initial cropping to cover important region. Then, a set of cropping candidates (the yellow rectangles in Figure 1 (b)) are generated around the attention box and the one with the highest aesthetics value is selected as the final cropping (Figure 1 (c)).

1.2 Contribution

Compared with previous arts, we treat the photo cropping task in a more natural and efficient way, with the following major contributions:

(1) A deep learning framework to combine attention and aesthetics components for photo cropping. We model photo cropping with a determining-adjusting process, where attention-guided cropping candidates generation is followed by aesthetics-aware cropping window selection, as shown in Figure 1 (e). Both tasks are achieved via a unified deep learning model, where attention information is exploited to avoid discarding important information, while the aesthetics assessment is employed for ensuring the high aesthetic value of the cropping result. The deep learning model is extended from the fully convolutional neural network, which naturally supports input images of arbitrary sizes, thus avoiding undesired deformation for evaluating aesthetic quality.

(2) High computation efficiency. Three ingredients are introduced in our approach for enhancing computational efficiency. First, instead of exhaustively searching all sub-windows in the sliding window fashion (e.g. [6]), our approach directly regresses the attention box and generates far less cropping candidates around the visually important areas. Second, the sub-networks for attention box prediction and for aesthetics assessment share several initial convolutional layers, and thus largely boost the efficiency by reusing the computation in these layers. Third, inheriting the advantage of recent object detection algorithms [7], [8], [9], our algorithm is trained to share convolutional features among cropping candidates. Regardless of the number of cropping candidates, these convolutional layers are calculated only once over the entire image, thus avoiding applying the network to each cropping candidate for repeatedly computing features. All these techniques help our approach to achieve a run time speed of 5 *fps*, significantly faster than previous solutions.

(3) Learning without cropping annotation. Use of deep leaning for vision problems typically requests a large amount of training data, which, for photo cropping, means a large amount of manually annotated cropping results. Such request is however very challenging, since photo cropping is very time consuming, and more importantly, is very subjective since it is difficult to offer a clear answer to what is a "groundtruth" cropping. Thus, training a network to directly output a cropping window is difficult and practically infeasible. We bypass this issue to use rich public data for human gaze prediction and photo aesthetics assessment. It is worth noting that, despite the absence of photo cropping data for training, our approach has shown great performance on the cropping task as shown in our thorough experiments.

These contributions together bring both effectiveness and efficiency to our proposed photo cropping algorithm. As described in Section 4, the thorough evaluations on popular benchmarks show clearly the advantage of our algorithm in comparison with state-of-the-art solutions.

2 Related Work

In this section, we first summarize representative works in visual attention prediction and aesthetics assessment (Section 2.1 and 2.2), respectively. Then, in Section 2.3, we give an overview of related works in photo cropping.

2.1 Visual Attention Prediction

Visual attention prediction is a classic computer vision problem that aims to predict scene locations where a human observer may fixate. This task, sometimes referred as *eye fixation prediction* or *visual saliency detection*, is for simulating human's ability of selectively paying attention to parts of the image instead of processing the whole scene in its entirety. A large amount of research effort has been devoted to this topic with many applications, such as image recognition [11], object segmentation [12], [13], [14], [15], image cropping [6], [16], etc. The output of attention prediction algorithms is usually a saliency map indicating the visual importance of each pixel.

Early visual attention models [17], [18] in the vision community are inspired by the studies in visual psychology and psychophysics of human attention. Those models can be further broadly classified into *bottom-up* approaches and *top-down* ones. Most of early models are based on the *bottom-up* mechanism, which is stimulus-driven and estimate human attention based on visual stimuli themselves without the knowledge of the image semantics. Such models [18], [19], [20], [21], [22] typically generate saliency cues based on various low-level features (*e.g.*, color, intensity, orientation) and heuristics (*e.g.*, center-surround contrast [23]) on limited human knowledge of visual attention, and combine them at multiple scales to create the final saliency map. By contrast to the bottom-up task-independent models, some *top-down* task-driven approaches [24], [25] are proposed that explore explicitly the understanding of the scene or task context. These approaches employ high-level features, such as person or face detectors learned from specific computer vision tasks. We refer readers to two recent surveys [26], [27] for more details of early attention models.

Deep learning-based attention models [28], [29], [31], [30], [32], [33] become increasingly popular in recent years, driven by the success of deep learning in object recognition and large-scale visual attention dataset (*e.g.*, SALICON [32]). Most of these models are variants of the *fully convolutional network* and generally produce more impressive results than non-deep learning competitors.

Traditional visual attention models concentrate on encouraging the consistency between the distribution of the predicted saliency and that of the real human fixations. Differently, in our approach, we are concerned more on predicting an attention bounding box, which covers the most informative regions of the image.

Another related topic in parallel is *salient object detection* [34], [39], [36], which can be dated to [35], [37] and has been extensively studied in computer vision in the past decade. Different from visual attention prediction, salient object detection specially focuses on detecting and uniformly highlighting one (multiple) salient object(s) in its (their) entirety. However, as stated in many literatures [38], [40], unlike fixation datasets, most salient object detection datasets are heavily biased to few objects. Therefore, for the sake of generalization capability and applicability, we choose visual attention prediction for photo cropping and use corresponding datasets (e.g., [32]), instead of the datasets of salient object detection.

2.2 Image Aesthetics Assessment

The main goal of aesthetics assessment is to imitate human's interpretation of the beauty of natural images. Many methods have been proposed for this topic, as surveyed in [41]. Traditionally, aesthetic quality analysis is viewed as a binary classification problem of predicting high- or low- quality of an image, or a regression problem of producing aesthetics scores. A common pipeline is to first extract visual features and then employ various machine learning algorithms to predict photo aesthetic values.

Early methods are mainly concerned on manually designing good feature extractors, which require a considerable amount of engineering skills and domain expertise. Some works [42], [43], [44], [45], [46] use hand-crafted aesthetics features according to photographic rules or experiences, such features include lighting, contrast, global image layout (rule-of-thirds), visual balance, typical objects (human, animals, plants), etc. These rule-based approaches are intuitive in that they explicitly model the criteria used by humans in evaluating the aesthetic quality of photos. Instead of using hand-crafted features, another option [47], [48] for image quality assessment is to leverage more generic image descriptors, such as the Fisher vector and bag of visual words, which are previously designed for image classification but also capable of capturing aesthetic properties.

More recently, **deep learning-based solutions** [49], [50], [51], [52], [53] have shown that image aesthetics representation may be better learned in an end-to-end data-driven manner. This trend is more and more popular with the growth of available training data, *i.e.*, from hundreds of images to millions of images. Such deep learning-based methods have greatly advanced the frontier of this topic.

2.3 Photo Cropping

Photo cropping is an important operation for improving visual quality of digital photos. Many methods have been proposed towards automating this task, and they can be roughly categorized into *attention-based* or *aesthetics-based*.

Attention-based approaches [1], [2], [3], [54] focus on preserving the main subject or visually important area in the scene after cropping. These methods usually choose the cropping window according to certain attention scores or object-level saliency map. These methods are usually good for removing unimportant content of an image, while sometimes fail to produce visually pleasant results due to the lack of consideration in image aesthetics.

Aesthetics-based approaches, by contrast, emphasize the general attractiveness of the cropped image. Those approaches [4], [5], [55], [56] are centered on composition-related image properties and low-level image features. Taking various aesthetical factors into account, they attempt to find the cropping candidate with the highest quality score. These methods are in favor of preserving visually attractive solutions, while at the risk of missing important area and generally suffer from expensive computation due to the need of evaluating a large amount of cropping candidates.

In general, conventional cropping methods search the region with the highest attention/aesthetics score in a number of candidate cropping windows. In this paper, we consider both attention and aesthetics information, and treat photo cropping as a cascade of first generating cropping candidates, via attention box prediction, and then selecting the best cropping window, via the aesthetics criteria. Our method shares the spirit of recent object detection algorithms [7], [8], [9]. In fact, a branch of our network learns to predict the bounding box covers visually important area, while the other branch estimates aesthetic value.

This paper extends a preliminary version appears in ICCV 2017 [57]. The improvements are multiple folds. First, we give a deeper insight into the proposed determining-adjusting based cropping protocol, with the comparison of previous sliding-judging strategy. This brings a new view into the rationale behind photo cropping. Second, we extend our attention box prediction network with supervised attention mechanism, outlining a complete model for better capturing the visual importance of input image and generating more accurate attention box prediction. It also improves the interpretability of our model and leads to an implicit deep supervision. Third, we offer a more in depth discussion of the proposed algorithm, including motivations, network structures and implementation. Forth, extensive experiments and user studies are conducted for thoroughly and insightfully examination. Last but not least, based on our experiments, we draw several important conclusions, which are expected to inspire future works in this direction.

3 Deep Learning-based Photo Cropping

We model photo cropping in a determining-adjusting framework, which first creates an initial cropping as a bounding box covering the most visually important area (attention-aware determining), and then selects the best cropping with the highest aesthetic quality from cropping candidates generated around the initial cropping (aesthetic-based adjusting). The cropping algorithm is decomposed into two cascaded stages, namely, attention-aware cropping candidates generation (Section 3.1) and aesthetics-based cropping selection (Section 3.2). A deep learning framework is thus designed with two sub-networks: an *Attention Box Prediction* (ABP) network and an *Aesthetics Assessment* (AA) network. Specifically, the ABP network is responsible for inferring

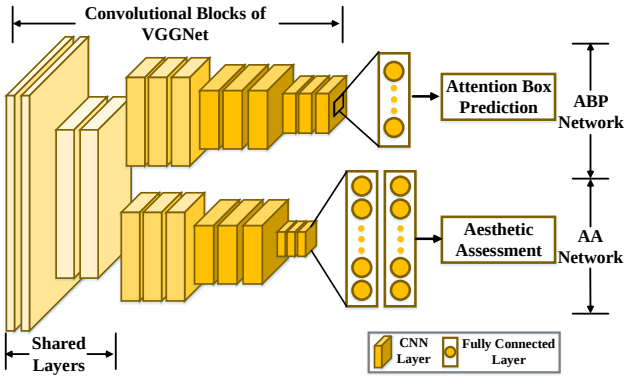


Fig. 2. Architecture of our deep cropping model. It consists of two sub-networks: Attention Box Prediction (ABP) network and Aesthetics Assessment (AA) network, which share several convolutional layers at the bottom.

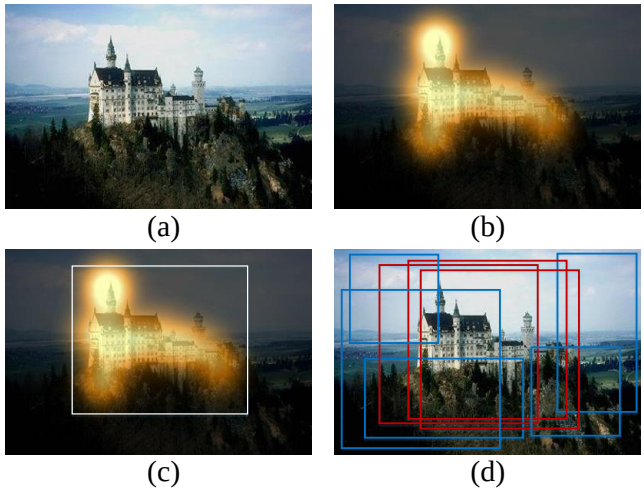


Fig. 3. (a) Input image I. (b) Ground truth attention map G. (c) Ground truth attention box generated via [3]. (d) Positive (red) and negative (blue) defaults boxes are generated for training ABP network according to ground truth attention box.

the initial cropping; and the AA network determines the final cropping. As demonstrated in Figure 2, these two networks share several convolutional blocks in the bottom and are based on fully convolutional network, which will be detailed in following sections. Finally, in Section 3.3, we give more details of our model in training and testing.

3.1 Attention-aware Cropping Candidates

In this section, we introduce our method for cropping candidates generation, which is based on an Attention Box Prediction (ABP) network. This network takes an image of any size as input and outputs a set of rectangular cropping windows, each with a score that stands for the prediction accuracy. Then the initial cropping is identified as the most accurate one, and various cropping candidates with different sizes and ratios are generated around it. After that, the final cropping is selected from those candidates according to their aesthetic quality based on an Aesthetics Assessment (AA) network (Section 3.2).

The initial cropping can be viewed as a rectangle that preserves the most informative part of the image while has minimum area. Searching for an optimal solution is common for attention-based cropping methods. Let $G \in [0,1]^{w \times h}$ be an attention mask of image I of size $w \times h$, and larger values in G indicate higher visual importance of corresponding pixels in I. Formally, we derive a set of cropping windows $\mathfrak{W}$ considering their importance or informativeness:

$$\mathfrak{W} = \Big\{ W \Big| \sum_{x \in W} G(x) \; > \; \lambda \sum_{x \in \{1..w\} \times \{1..h\}} G(x) \Big\}, \tag{1}$$

where $\lambda \in [0,1]$ is a threshold. Then the optimum cropping rectangle $\widehat{W}$ is defined as the one with minimum area:

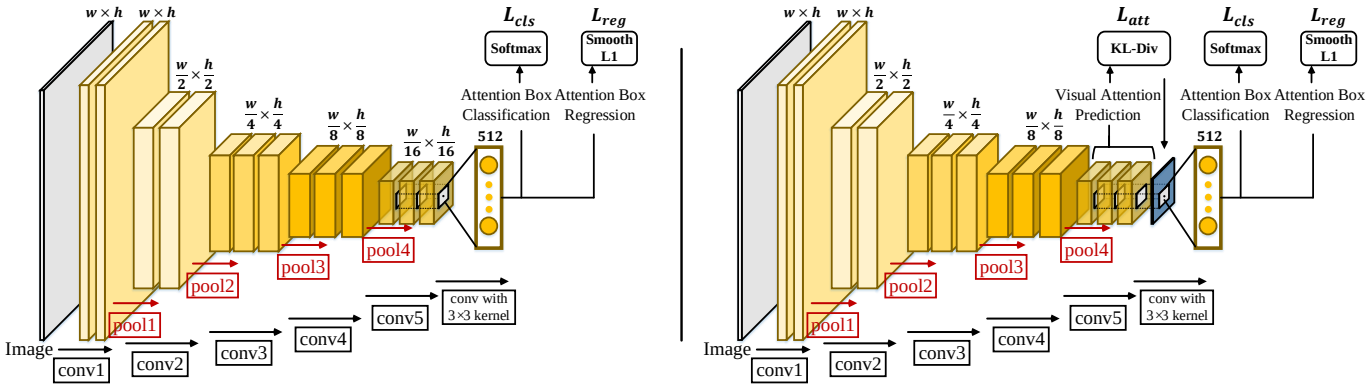
$$\widehat{W} = \underset{W \in \mathfrak{W}}{\operatorname{argmin}} |W|. \tag{2}$$

Equ. 2 can be solved via sliding window searching with $\mathcal{O}(w^2h^2)$ computation complexity, while a recent method [3] shows it can be solved with computation complexity of $\mathcal{O}(wh^2)$ (assuming $h < w$).

Different from the above time consuming strategy, we design a neural network for predicting an optimal attention box. Given a training sample (I, G) consisting of an image I of size $w \times h \times 3$ (Figure 3 (a)), and a groundtruth attention map $G \in [0,1]^{w \times h}$ (Figure 3 (b)), the optimum rectangle $\widehat{W}$ defined in Equ. 2 is treated as the groundtruth attention prediction box. Here we apply the method in [3] for generating $\widehat{W}$ over G (Figure 3 (c)) for computation efficiency, and set $\lambda = 0.9$ for preserving most informative areas. Then the task of attention box prediction can be achieved via bounding box regression similar as in object detection [7], [8], [9]. Note that, our ABP network is not limited to specific attention scores, and other attention models can be used for generating groundtruth bounding box as well.

Figure 4 illustrates the architecture of the ABP network. The bottom of this network is a stack of convolutional layers, which are borrowed from the first five convolutional blocks of VGGNet [58]. In our conference version [57], we build the bounding box regression layers upon the last convolutional layer with a small network of a 3×3 kernel (see Figure 4 (a)). Thus the network is trained to directly produce the attention box estimation.

We further improve the ABP network with extra supervision from the visual attention map G. It is demonstrated in Figure 4 (b), showing in the last convolutional layer. Specifically, we first generate an intermediate output Y for predicting the visual attention map (the blue cuboid in Figure 4 (b)) by a convolution layer with a 1×1 kernel and *sigmoid* activation. Then the attention map Y is concatenated with the last convolutional layer in the channel direction, and the merged feature maps are fed into the bounding box prediction layers for generating the final attention box. Such design is based on the observation that attention box is derived from the visual attention via Equ. 2. The visual attention can act as a strong prior for attention box and teach the network to infer the attention box via leveraging the strong relevance between visual attention and attention box. More specially, given the resized groundtruth attention map $G' \in [0,1]^{\frac{w}{16} \times \frac{h}{16}}$ and the corresponding prediction map



(a) ABP network used in our conference paper [57]

(b) Improved ABP network

Fig. 4. Architecture of the Attention Box Prediction (ABP) network, where the blue cuboid in (b) indicates the predicted attention map.

$Y \in [0,1]^{\frac{w}{16}\times\frac{h}{16}}$, we adopt the *Kullback-Leibler Divergence* (KL-Div) for measuring the training loss:

$$\mathcal{L}_{\text{att}}(Y, G') = \sum\nolimits_i g_i \log\left(\frac{g_i}{y_i}\right). \tag{3}$$

The KL-Div measure, whose minimization is equivalent to cross-entropy minimization, is widely used to learn visual attention models in deep networks.

Then we slide a small network of a 3×3 kernel and 512 channels over the merged feature map, thus generating a 512-dimensional feature vector for each sliding location. The feature vector is further fed into two fully-connected layers: a box-regression layer for predicting attention bounding box and a box-classification layer for determining whether a box belongs to attention box. For a given location, those two fully-connected layers predict box offsets and scores over a set of default bounding boxes, which are similar to the *anchor boxes* used in Faster R-CNN [8].

To train the ABP network for bounding box prediction, we need to decide the positive and negative training boxes (samples) correspond to the groundthe attention box and train the network accordingly. We treat a box as a positive box if it has the Intersection-over-Union (IoU) score with the groundtruth box of larger than 0.7, or it has the largest IoU score. In such case, we give it a positive label ($c = \mathbf{1}$). By contrast, we treat a box as negative ($c = \mathbf{0}$) if it has an IoU score lower than 0.3 and drop other default boxes. The above process is illustrated in Figure 3(d). For the preserved boxes, we define $\bar{p}_i^c \in \{1, 0\}$ as an indicator for the label of the i-th box and vector $\bar{t}$ as a four-parameter coordinate (coordinates of center, width and height) of the groundtruth attention box. Similarly, we define p_i^c and t_i as predicted confidence over c class and predicted attention box of the i-th default box. With the above definition, we consider the following loss function for bounding box prediction, which is derived from object detection [59], [8], [60]:

$$\mathcal{L}_{\text{box}}(p, t) = \sum\nolimits_i \mathcal{L}_{\text{cls}}(p_i, \bar{p}_i) + \sum\nolimits_i \bar{p}_i^{\mathbf{1}} \mathcal{L}_{\text{reg}}(t_i, \bar{t}). \tag{4}$$

The classification loss $\mathcal{L}_{\text{cls}}$ is the softmax loss over confidences of two classes (attention box or not). The regression loss $\mathcal{L}_{\text{reg}}$ is a smooth L1 loss [59] between the predicated box and the ground truth attention box, and it is only activated for positive default boxes.

With the above definition, the ABP network is trained via minimizing the following overall loss function:

$$\mathcal{L} = \mathcal{L}_{\text{att}} + \mathcal{L}_{\text{box}}, \tag{5}$$

where $\mathcal{L}_{\text{att}}$ (defined in Equ. 3) is an intermediate loss for directly feeding supervision into the hidden layers, and the learned attention acts as a strong prior to improve the final bounding box prediction. The terminal loss $\mathcal{L}_{\text{box}}$ (defined in Equ. 4) is for regressing the bounding box location and predicting the attention box score.

Trained on existing attention prediction datasets, the ABP network learns to generate reliable attention boxes. Then we select the one with the highest prediction score ($p_i^{\mathbf{1}}$) as the initial cropping. This initial cropping covers the most informative part of the image, and it simulates human's placement of a cropping window around the desired area (Figure 5(a)). Next, we generate a set of cropping candidates around the initial cropping, simulating human's adjusting of the location, size and ratio of the initial cropping. A rectangle can be uniquely determined via the coordinates of its top-left and bottom-right corners. For the top-left corner of the initial cropping, we define a set of offsets $\{-40, -32, \cdots, -8, 0\}$ in both x- and y-axes. Similarly, a set of offsets $\{0, 8, ..., 32, 40\}$ in x- and y-axes are defined for the bottom-right corner. By disturbing the top-left and bottom-right corners with these offsets,[1] we generate $6^4 = 1,296$ cropping candidates in total, which is far less than the sliding windows needed by traditional exhaustive cropping methods. Each of cropping candidates is designed to cover the entire initial cropping area, since the initial cropping is a minimum importance-preserving rectangle to be maintained during the cropping process (Figure 5(b)).

3.2 Aesthetics-based Cropping Window Selection

With the attention-aware cropping candidates by the ABP network, we select the most aesthetically-pleasant one as the final cropping. It is important to consider aesthetics for photo cropping, since beyond preserving the important content, a good cropping should also deliver pleasant viewing experience. For analyzing the aesthetic quality of each cropping candidates, one choice is to train

1. Since we resize the input image with $\min(w, h) = 224$, we find the largest offset of 40 to be sufficient.

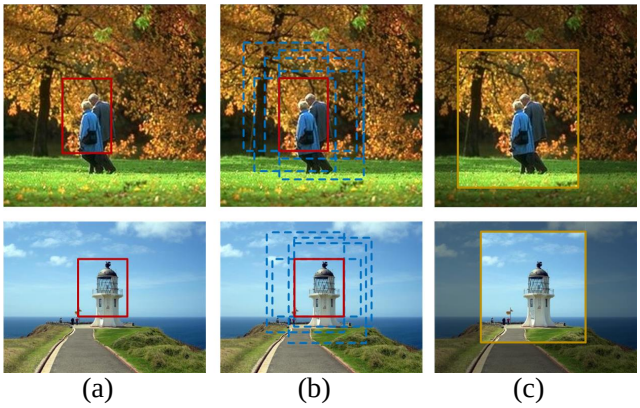


Fig. 5. (a) Initial cropping (red rectangle) predicted by the ABP network. (b) Cropping candidates (blue rectangles) generated around the initial cropping. (c) The final cropping selected as the candidate with the highest aesthetic score by the AA network.

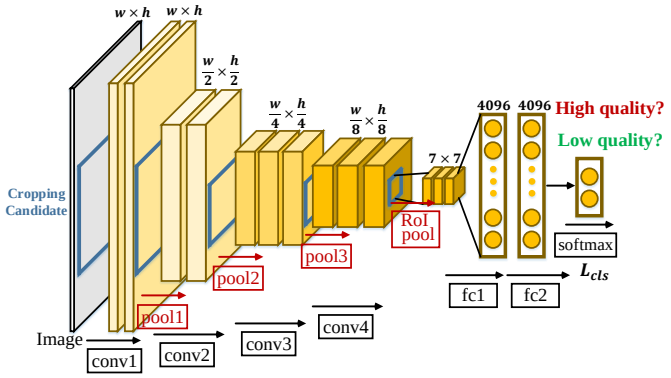


Fig. 6. Architecture of the Aesthetics Assessment (AA) network.

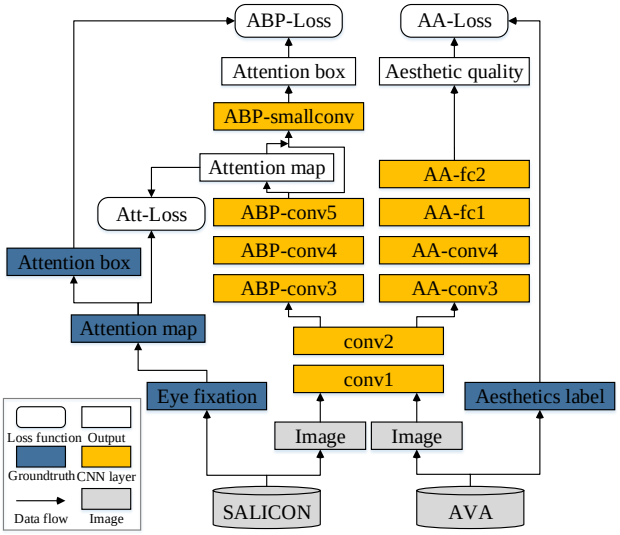


Fig. 7. Schematic diagram of our model in training.

an aesthetics assessment network, and iteratively applying forward-propagation for each cropping candidate over this network. This straightforward strategy is obviously very time-consuming. Inspired by the recent advantages of object detection, which shares convolutional features between regions, we propose to build a network that analyzes aesthetic values of all candidates simultaneously.

We achieve this via an Aesthetics Assessment (AA) network (Figure 6), which takes an entire image and a set of cropping candidates as input, and outputs the aesthetic values of the cropping candidates. The bottom of the AA network is the first four convolutional blocks of VGGNet [58] excluding the *pool4* layer. Here we adopt a relatively shallow network mainly due to two reasons. First, aesthetics assessment is a relatively easy problem (with only two labels: high quality *vs* low quality) compared with image classification (with 1000 classes for ImageNet). Second, for an image of size of $w \times h \times 3$, the spatial dimensions of the final convolutional feature map of AA network is $\frac{w}{8} \times \frac{h}{8}$, which preserves discriminability for the offsets defined in Section 3.1.

On the top of the last convolutional layer, we adopt a region of interest (RoI) pooling layer [8], which is a special case of spatial pyramid pooling (SPP) [7], to extract a fixed-length feature vector from the last convolutional feature map. The RoI pooling layer uses max-pooling to convert the features inside any cropping candidate into a small feature map with a fixed-dimensional vector, which is further fed into a sequence of fully-connected layers for aesthetic quality classification. This operation allows us to handle images with arbitrary aspect ratios, thus avoiding undesired deformation in aesthetics assessment. For a cropping candidate of size of $w' \times h'$, the RoI pooling layer divides it into $n \times n$ (n=7 in our experiments) spatial bins and applies max-pooling for the features within each bins.

For training, given an image from existing aesthetics assessment datasets, it takes an aesthetic label $c \in \{\mathbf{1}, \mathbf{0}\}$, where $\mathbf{1}$ indicates high aesthetic quality and $\mathbf{0}$ indicates low quality. We resize the image so that $\min(w, h) = 224$, same as for the ABP net, and the whole image can be viewed as a cropping candidate for training. For the i-th training image, we define $\bar{q}_i^c \in \{1, 0\}$ as an indicator for its aesthetics-quality label and q_i^c as its predicted aesthetics-quality score for class c.

Based on the above definition, the training of the AA network is done by minimizing the following softmax loss over N training samples:

$$\mathcal{L}_{\text{cls}}(q, \bar{q}) = -\frac{1}{N} \sum\nolimits_i \sum\nolimits_{c \in \{\mathbf{1},\mathbf{0}\}} \bar{q}_i^c \log(\hat{q}_i^c), \tag{6}$$

$$\hat{q}_i^c = \frac{\exp(q_i^c)}{\sum_{c' \in \{\mathbf{1},\mathbf{0}\}} \exp(q_i^{c'})}. \tag{7}$$

With the cropping candidates generated from the APB network, the AA network is capable of producing their aesthetics-quality scores ($\{q_i^{\mathbf{1}}\}_i$), where the one with the highest score is selected as the final cropping (Figure 5 (c)).

3.3 Implementation Details

3.3.1 *Training*

Two large-scale datasets: SALICON [32] and AVA [61], are used for training our model.

The SALICON dataset is used for training our ABP network. It contains 20,000 natural images with eye fixation

annotations. In the area of saliency prediction, the publication of SALICON dataset has enabled end-to-end training of deep architectures specifically for attention prediction. To obtain smooth saliency maps, we follow [32] to apply a Gaussian filter with a small kernel for filtering a binary mouse-clicking map into a grey-scale human attention map. To obtain the groundtruth attention box, we apply the algorithm in [3] to the saliency map to generate attention bounding boxs according to Equ. 2 with $\lambda = 0.9$.

The AVA dataset is the largest publicly available aesthetics assessment benchmark, containing about 250,000 images in total. The aesthetics quality of each image was rated on average by roughly 200 people with the ratings ranging from one to ten, with ten indicating the highest aesthetics quality. Followed the work in [49], [51], [53], [61], about 230,000 images are used for training our AA network. More specifically, images with mean ratings smaller than 5 are assigned as low quality and the rest as high quality. More details of the two datasets can be found in Section 4.1.

Our two sub-networks are trained simultaneously. In each training iteration, we use a min-batch of 10 images, 5 of which are from the SALICON dataset with the groundtruth attention boxes and the rest from the AVA dataset with aesthetics quality groundtruth. Before feeding the input images and ground-truth to the network, we scale the images such that the short side is of size 224. The whole training scheme of our model is presented in Figure 7. The $conv1$ and $conv2$ blocks are shared between both the tasks of attention box prediction and esthetics assessment, and they are trained for both the tasks simultaneously using all the images in the batch. For the layers specialized for each sub-network, they are trained using only those images in the batch with the corresponding ground-truth.

Both ABP and AA networks are initialized from the weights of VGGNet [58], which is pre-trained on the large-scale image classification dataset, ImageNet [62], with 1M images. Our model is implemented with Keras and trained with the Adam optimizer [63]. The learning rate is set to 0.0001. The networks were trained over 10 epochs. The entire training procedure takes about 1 day with an Nvidia TITAN X GPU.

3.3.2 Testing

While our two sub-networks are trained in parallel, they work in a cascaded way (see Figure 8) during testing. Given an input image (resized such that $\min(w, h) = 224$) for cropping, we first gain a set of attention boxes generated by forward propagation on the APB network. Then the initial cropping is selected as the one with the highest accuracy of attention box prediction. After that, a set of cropping candidates are generated around the initial one. Since the two initial convolutional blocks are shared between the ABP and the AA networks, we directly feed the cropping candidates and the convolutional feature of last layer of $conv2$ into the AA network. The final cropping is selected as the cropping candidate with best aesthetic quality. The whole algorithm runs at about 5 fps.

4 Experimental Results

In this section, we first detail the datasets used for training and testing in Section 4.1. Then we examine the performance

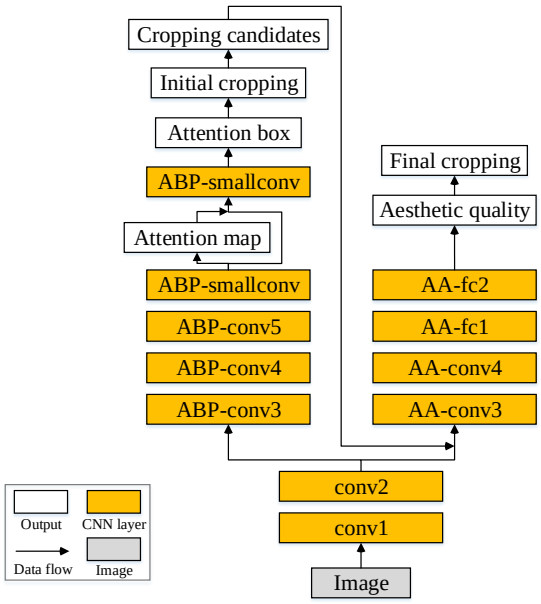


Fig. 8. Schematic diagram of our model in testing.

of our ABP and AA networks on their specific tasks (Section 4.2 and 4.3). The goal of these experiments is to investigate the effectiveness of individual components instead of comparing them with the state-of-the-arts. Then, in Section 4.4, we evaluate the performance of our whole cropping model on two widely used photo cropping datasets with other competitors. In Section 4.5, detailed discussions for limitation and future work are presented.

4.1 Datasets

There are totally six datasets, namely SALICON [32], PASCAL-S [38], AVA [61], Image Cropping Dataset from MSR (MSR-ICD) [5], FLMS [6], and Flickr Cropping Dataset (FCD) [10], used in our experiments. Some statistics of these datasets and experimental settings are summarized in Table 1. SALICON and PASCAL-S are employed, respectively, for training and testing our ABP network (Section 4.2); AVA is used for training and testing our AA network (Section 4.3); MSR-ICD, FLMS and FCD are used for accessing the performance of our full cropping solution (Section 4.4). Next we give detailed descriptions for each of the datasets.

- **SALICON**. This is one of the largest saliency datasets available in the public domain. It contains 20,000 natural images from the MSCOCO dataset [64] with eye fixation annotations that are simulated through mouse movements of users on blurred images. These images contain diverse indoor and outdoor scenes and display a range of scene clutter. 10,000 images are marked for training, 5,000 for validation and 5,000 for testing. We use the training and validation sets (with publicly available annotations) for training our ABP network. Since the fixation data for the test set is held-out, we turn to another widely used dataset, PASCAL [38], for accessing the performance of ABP network.
- **PASCAL-S**. This dataset contains 850 natural images from the validation set of PASCAL 2010 [65] segmentation challenge. There are totally eight subjects are instructed to perform the "free-viewing" task in the fixation experiment.

TABLE 1
Datasets used for training and testing our cropping model.

	Dataset	Ref	Year	#Images	Purpose	
					Train	Test
ABP network	SALICON	[32]	2015	20,000	✓	
	PASCAL-S	[38]	2014	850		✓
AA network	AVA	[61]	2012	~250,000	✓	✓
Deep-cropping	MSR-ICD	[5]	2013	950		✓
	FLMS	[6]	2014	500		✓
	FCD	[10]	2017	1,743		✓

Each image is presented for 2 seconds, and the eye gaze data is recorded using Eyelink 1000 eye-tracker, at 125Hz sampling rate. The smooth attention map for each image is obtained by blurring the fixation map with a fixed Gaussian kernel ($\sigma = 0.005$ of the image width).

• **AVA**. The Aesthetic Visual Analysis (AVA) dataset contains about 250,000 images in total. These images are obtained from DPChallenge.com and labeled for aesthetic scores. Specifically, each image receives 78~549 votes of a score ranging from 1 to 10. For the task of binary aesthetic quality classification, images with an average score higher than 5 are treated as positive examples, and the rest image are treated as negative ones. Accordingly, a large-scale standardized partition is obtained, which has about 230,000 images for training and about 20,000 images for testing.

• **MSR-ICD**. The MSR-ICD dataset includes 950 images which are originally from an image aesthetics assessment database [46]. The photos are acquired from the professional photography websites and contributed by amateur photographers and span a variety of image categories, including animal, architecture, human, landscape, night, plant and man-made objects. Each image is carefully cropped by three expert photographers.

• **FLMS**. The FLMS dataset contains 500 natural images collected from Flickr. For each image, 10 expert users on Amazon Mechanical Turk who passed a strict qualification test are employed for cropping groundtruth box.

• **FCD**. It consists of 1,743 images collected from Flickr. Seven workers on Amazon Mechanical Turk were recruited for annotation. The images are split into a training set of 1,369 images and a test set of 374 images.

4.2 Performance of the ABP Network

We first evaluate the ABP network on the PASCAL-S dataset [38], which is widely used for attention prediction. With the binary eye fixation images, we follow [38] to generate gray-scale attention maps. Then, as described in Section 3.3, we generate a groundtruth attention box for each image.

TABLE 2
Attention box prediction with IoU for PASCAL-S dataset [38].

Method	Ours [57]	ITTI [17]	AIM [18]	GBVS [19]	SUN [20]
IoU	0.517	0.318	0.327	0.319	0.273
Method	**Ours**	DVA [21]	SIG [66]	CAS [67]	SalNet [33]
IoU	**0.583**	0.346	0.272	0.356	0.379

We test eight state-of-the-art attention models including ITTI [17], AIM [18], GBVS [19], SUN [20], DVA [21], SIG [66], CAS [67] and SalNet [33]. Previous attention models are for imitating human visual attention behavior, and their output is a continuous saliency map. In contrast, our AA network generates an important bounding box as an initial cropping. Thus PR curves or AUC curves used in visual attention prediction cannot be directly applied for comparison. For the sake of a relatively fair comparison, we first extract the attention boxes of above methods via the same strategy used for generating the groundtruth bounding box. Then we apply the Intersection over Union (IoU) score for quantifying the quality of extracted attention boxes. We also report the results from our preliminary conference version [57].

The quantitative results are illustrated in Table 2. As seen, our attention box prediction results are more accurate than previous attention models, since our ABP network is specially designed for this task. Additionally, comparing the performance of our conference version, the improvement is significant (0.517→0.583). This is mainly due to the incorporation of visual attention supervision in our ABP network, which offers strong prior knowledge for attention box prediction.

TABLE 3
Aesthetics assessment accuracy on the AVA dataset [61].

Method	**Ours**	AVA [61]	RAP-DCNN [49]	RAP-RDCNN [49]
Accuracy	**0.770**	0.667	0.732	0.745
Method	**Ours**	RAP2 [68]	DMA-SPP [51]	DMA [51]
Accuracy	**0.770**	0.754	0.728	0.745
Method	**Ours**	DMA-Alex [51]	ARC [52]	CPD[53]
Accuracy	**0.770**	0.754	0.773	0.774

4.3 Performance of the AA Network

We adopt the testing set of the AVA dataset [61], as described in Section 3.3, for evaluating the performance of our AA network. The testing set of AVA dataset contains 19,930 images. The testing images with mean ratings smaller than 5 are labeled as low quality; otherwise they are labeled as high quality.

We compare our methods with the state-of-the-art methods including AVA [61], RAP [49], RAP2 [68], DMA [51], ARC [52] and CPD [53], where AVA offers the state-of-the-art result based on manually designed features while other methods are based on deep learning model.

We opt the overall accuracy metric, which is the most popular evaluation criterion in the research area of image aesthetics assessment, for quantitative evaluation. It can be expressed as Accuracy $= \frac{TP+TN}{P+N}$, where TP, TN, P and N refer to true positive, true negative, total positive, and total negative, respectively. This metric accounts for the proportion of correctly classified samples.

It is clear from Table 3 that, our AA network achieves state-of-the-art performance even with a relatively simple network architecture. In Figure 9, we present some examples with aesthetics values predicted by our AA network.

Overall, our two sub-networks generate the promising results aligned with existing top-performance approaches. This is mainly due to a relatively shallow network and simple network architecture, compared with exiting deep

(a) Images with highest aesthetics values predicted by our AA network

(b) Images with lowest aesthetics values predicted by our AA network

(c) Images classified as high-quality but labeled as low-quality

(d) Images classified as low-quality but labeled as high-quality

Fig. 9. Aesthetics assessment results via our AA network. The images with the highest predicted aesthetics values and those with the lowest predicted aesthetics values are presented in (a) and (b), respectively. (c) and (d) show the images that are miscategorized.

TABLE 4
Performance of automatic image cropping on MSR-ICD dataset [5]. Higher IoU score and lower BDE indicate better cropping predictor.

Method	* Photographer 1		Photographer 2		Photographer 3		Average	
	IoU ↑	BDE ↓	IoU ↑	BDE ↓	IoU ↑	BDE ↓	IoU ↑	BDE ↓
ATC [16]	0.605	0.108	0.628	0.100	0.641	0.095	0.625	0.101
AIC [3]	0.469	0.142	0.494	0.131	0.512	0.123	0.491	0.132
LCC [5]	0.748	0.066	0.728	0.072	0.732	0.071	0.736	0.0670
MPC [69]	0.603	0.106	0.582	0.112	0.608	0.110	0.598	0.109
SPC [4]	0.396	0.177	0.394	0.178	0.385	0.182	0.391	0.179
ARC [52]	0.448	0.163	0.437	0.168	0.440	0.165	0.442	0.165
Ours [57] (conference version)	0.813	**0.030**	0.806	0.032	0.816	0.032	0.812	0.031
Ours	**0.815**	0.031	**0.810**	**0.030**	**0.830**	**0.029**	**0.818**	**0.029**

* MSR-ICD dataset offers separate annotations from three different expert photographers.

TABLE 5
Performance of automatic image cropping on the FLMS dataset [6]. Higher IoU score and lower BDE indicate better cropping results.

Dataset	Method	Measure	
		IoU ↑	BDE ↓
FLMS	ATC [16]	0.72	0.063
	AIC [3]	0.64	0.075
	LCC [5]	0.63	–
	MPC [69]	0.41	–
	VBC [6]	0.74	–
	Ours [57] (conference version)	0.81	0.057
	Ours	**0.83**	**0.052**

- The authors in LCC [5], MPC [69] and VBC [6] have not released results with the BDE measure.

TABLE 6
Performance of automatic image cropping on the test set of FCD [10]. Higher IoU score and lower BDE indicate better cropping results.

Dataset	Method	Measure	
		IoU ↑	BDE ↓
FCD	ATC [16]	0.58	0.10
	AIC [3]	0.47	0.13
	ATC [16] + eDN [29] (MaxAvg)	0.35	0.17
	ATC [16] + eDN [29] (MaxDiff)	0.48	0.13
	ATC [16] + BMS [70] (MaxAvg)	0.34	0.18
	ATC [16] + BMS [70] (MaxDiff)	0.39	0.16
	SVM+DeCAF$_7$	0.51	0.13
	AVA+DeCAF$_7$	0.52	0.12
	FCD+DeCAF$_7$	0.60	0.10
	Ours [57] (conference version)	0.63	0.09
	Ours	**0.65**	**0.08**

learning aesthetics network. Considering the shared convolutional layers in the bottom of these two networks, our model achieves a good tradeoff between performance and computation efficiency. More importantly, the robustness of these two basic components greatly contributes to the high-quality of our cropping suggestions, which will be detailed in the next section.

4.4 Performance of Cropping Network

4.4.1 Evaluation on MSR-ICD and FLMS Datsets

We first evaluate our full cropping model on two widely-used image cropping datasets, including the Image Cropping Dataset from MSR (MSR-ICD) [5] and the FLMS dataset [6] . We adopt the same evaluation metrics as [5], *i.e.*, the IoU score and the Boundary Displacement Error (BDE) to measure the cropping accuracy of image croppers. BDE is defined as the average displacement of four edges between the cropping box and the groundtruth rectangle:

$$\text{BDE} = \sum\nolimits_i ||B_i^g - B_i^c||/4, \tag{8}$$

where $i \in \{left, right, bottom, up\}$ and $\{B_i\}_i$ denote the four edges of the cropped window or groundtruth cropping. Note that BDE has to be normalized by the width or height of the image. Clearly, a good cropping solution favors a high IoU score and a low BDE.

We compare our cropping method with two main categories of image cropping methods, *i.e.*, *attention-based* and *aesthetics-based* methods.

For attention-based methods, we select the ATC algorithm [16] which is a classical image thumbnail cropping method, and the AIC algorithm [3]. The results of AIC algorithm are obtained via applying cropping window researching method [3] with top-performing saliency detection method. Here we apply context-aware saliency [67] and optimal parameters, as suggested by [3], for maximizing its performance. For aesthetics-based methods, we select LCC [5], MPC [69], and VBC [6]. In addition, we consider SPC, which is an advanced version of [4], as described in [5]. Additionally, we adopt a recent aesthetics ranking method [52] combined with sliding window strategy as a baseline: ARC. We select the cropping as the one with the highest ranking score from sliding windows.

The quantitative comparison results on the MSR-ICD and FLMS datasets are demonstrated in Table 4 and Table 5, respectively. As seen, our cropping method achieves the best performance on both datasets. The improvement over our conference version verifies the effectiveness of our improved ABP network. Qualitative results on the MSR-ICD and FLMS datasets are presented in Figure 10.

4.4.2 Evaluation on FCD Datset

We further test the proposed cropping method on the test set of the recently released FCD [10] dataset. Following the settings in FCD dataset, we extend the *attention-based* ATC algorithm [16] with two state-of-the-art attention methods, *i.e.*, BMS [70] and eDN [29], using two search strategies, *i.e.*, MaxAvg (searching an optimal cropping window with the highest average saliency) and MaxDiff (maximizing the difference of average saliency between the crop and the outer region). For *aesthetics-based* methods, we consider three baselines in [10]: SVM+DECAF$_7$, AVA+DECAF$_7$ and FCD+DECAF$_7$, corresponding to a combination of the SVM classifier and DECAF$_7$ features [71], training on the AVA and FCD datasets, respectively. As summarized in Table 6, the results on the FCD dataset demonstrate again that our method compares favorably with the previous state-of-the-art methods using the two evaluation metrics.

4.4.3 User Study

Since photo cropping is a human-centric task, we conduct a user study for assessing the quality of cropping suggestions from our system and other competitors, including ATC [16] and AIC [3]. A corpus of 20 participants (8 females and 12 males) with diverse backgrounds and ages were recruited to participate in the user study. None of the participants had received any special technical instructions or had any prior knowledge about the experimental hypotheses. 200 images randomly selected from the MSR-ICD [5] and FLMS [6] datasets are used in this user study. The original image and its cropped versions from our method and other competitors are presented to the participants. Each participant examines all the selected images and is required to answer which cropped image they prefer. Figure 11 shows the distribution of votes averaged over all participants. As seen, our method receives the most overall votes, confirming the strong preference of the proposed method over other methods.

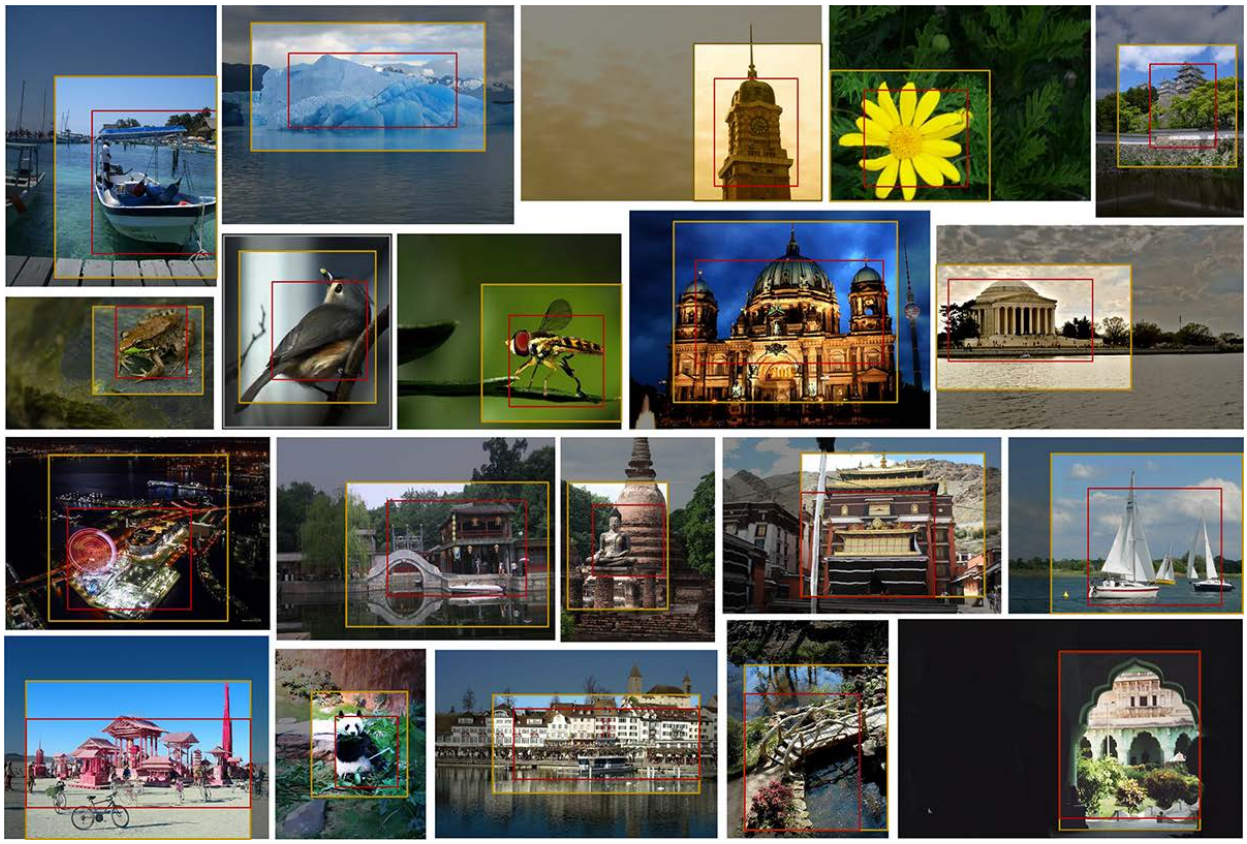


Fig. 10. Qualitative results on MSR-ICD [5] and FLMS [6] datasets. The red rectangles indicate the initial cropping generated by the ABP network, and the yellow windows correspond to the final cropping selected by the AA network.

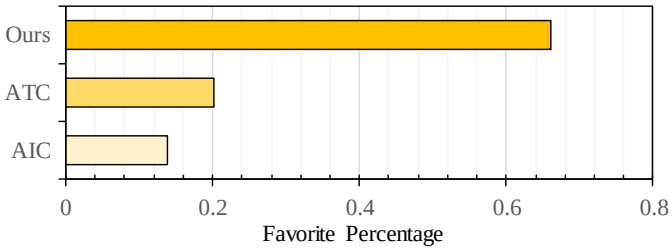


Fig. 11. User preference rate in user study.

4.4.4 Ablation Study

To give a deeper insight of the proposed cropping method, we study different ingredients and variants of our method. We experiment on the FLMS dataset [6] and measure the performance using the IoU metric. Four baselines derived from our method are considered:

• *ABP*: It directly uses the initial cropping from the ABP network as the final cropping.

• *Sliding window+AA*: We apply sliding windows (∼10,000 windows for an image with typical resolution of 224×224) and use the AA network to select the best aesthetics-preserved one as the final cropping.

• *ATC+AA*: It corresponds to the results that we treat the cropping results from attention-based cropping method ATC [16] as the initial cropping and further apply the AA network for determining the final cropping.

• *AIC+AA*: Similar to *ATC+AA*, we combine the AA network with the results from attention-based AIC [3] for outputting the final cropping.

TABLE 7
Ablation study on FLMS dataset [6].

Aspect	Description	Measure	
		IoU↑	Time(s)↓
full model	ABP+AA	0.83	0.23
variant	ABP	0.77	0.12
	Slidingwindow+AA	0.69	134
	ATC[16]+AA	0.78	1.3
	AIC[3]+AA	0.73	32.5
competitor	ATC[16]	0.72	1.1
	AIC[3]	0.64	32.3
cropping candidates	Randomly sampling (*1,000 candidates*)	0.80	0.23
	Larger sampling step (*step=16*)	0.81	0.23

The evaluation results and computation time are summarized in Table 7. We can draw the following three important conclusions:

1) **Aesthetics is important for photo cropping.** The improvement brought from AA network (ABP+AA: 0.83 *vs* ABP: 0.77, ATC+AA: 0.78 *vs* ATC: 0.72, AIC+AA: 0.73 *vs* AIC: 0.64) indicates that the cropping performance is benefited from aesthetic assess-

ment. This conclusion aligns with the claims shared by pervious aesthetics-based cropping methods.

2) **Both visual importance and photo aesthetics are critical.** A drop of performance can be observed when only considering photo aesthetics (ABP+AA: 0.83 *vs* Sliding window+AA: 0.78). This observation can be attributed to omitting important image content when only considering photo aesthetics. It supports one of our motivations to combine visual importance and photo aesthetics together for correctly determining the cropping.
3) **The proposed cropping solution achieves high computation efficiency.** With the full implementation, the proposed algorithm achieves a high processing speed of 5 fps on a modern GPU, which is faster than other competitors.

To study the influence of the sampling strategy of our cropping candidates (Section 3.1), we further evaluate the following two baselines:

• *Randomly sampling*: Instead of using a set of fixed size offsets, we randomly extract 1,000 cropping candidates, all of which cover the initial cropping area.

• *Larger sampling step*: We enlarge the original sampling steps (=8) as 16, then apply AA network for selecting the final cropping.

From Table 6, we can observe performance drops of these two baselines. For random sampling, since the feature map of AA network is with $\times 8$ downsampling, some similar neighbor candidates may be repeatedly considered while some other important candidates may be missed. When we increase the sampling step, the performance becomes worse since some candidate regions are ignored. Overall, the proposed cropping algorithm that combines the ABP and AA networks achieves the best performance and is much more computationally efficient.

4.5 Discussions

4.5.1 Limitations

The proposed algorithm suffers from a few limitations. A potential drawback of utilizing visual attention is that unfaithful importance maps might negatively affect cropping results. The attention model (the ABP network) may omit parts of a salient object which occupies a large portion of the scene (see examples in Figure 12). This issue can be partly alleviated by considering the aesthetics quality from the AA network. Besides, since most of the training images in the AVA dataset are manually selected and pre-manipulated, the discriminability of the AA network may be limited with daily raw images. For training the AA network, the negative samples and the positive samples are from different scenes. However, such image-level aesthetics annotation is insufficient to provide enough supervision for the AA network for rating cropping cases from the same original image. For remedying this, more negative training examples with false cropping (*e.g.*, splitting an important object into parts) should be mined for training a more robust AA model.

(a) (b)

Fig. 12. Two cases for illustrating the limitations of our cropping solution, where the left images in (a) (b) show the importance maps and initial cropping (red rectangles) generated by the ABP network, the yellow windows in the right images are the final cropping selected by the AA network. We can find that the ABP network tends to select the most informative but small parts, which may discard some parts of a large object. See Section 4.5.1 for more discussion.

4.5.2 Future work

The proposed method is among the first attempts to apply deep learning for photo cropping, and opens various research directions that are worth future exploration.

• **Bottom-up attention vs top-down attention:** Similar to most previous attention based cropping algorithms, our method employs bottom-up attention model to determine the image parts to preserve. The bottom-up model imitates the selective mechanisms of human visual system in general scenes without considering high-level information. It is interesting to explore the integration of top-down task-driven attention into the proposed cropping framework. Such attention may help reveal the rationale behind human cropping behavior, *e.g.*, understanding the searching strategy of human, examining the correlation between purely visual importance and cropping-specific importance.

• **Classification vs ranking:** In our current approach, we formulate aesthetics analysis as a binary classification problem (*i.e.*, low- or high-aesthetics). However, the aesthetics assessment may be more of a ranking problem, since individuals have different aesthetics tastes but are more consistent with the relative aesthetic ranks. This can be achieved by specially designed aesthetics rating networks and ranking loss (like [52]), thus our cropping model may be more powerful and consistent with human aesthetic preference among different cropping cases.

• **Incorporating high-level human knowledge:** In photo aesthetics analysis, numerous efforts have been seen in designing features for encapsulating the intertwined aesthetic rules. It might be promising to incorporate human-knowledge of photographic rules (*e.g.*, region composition and rule of thirds) into our current cropping solution, since such domain knowledge is still instructive and widely used in photographic practice and visual design.

5 Conclusions

In this work, we proposed a deep learning-based photo cropping approach with the determining-adjusting philosophy. The proposed deep model is composed of two sub-networks: an Attention Box Prediction (ABP) network and an Aesthetics Assessment (AA) network, both of which share multiple initial convolution layers. The ABP network infers initial cropping as a bounding box covering the visually important area (attention-aware determining), and then the AA network selects the best cropping with the

highest aesthetic quality from a few cropping candidates generated around the initial cropping (aesthetic-based adjusting). Extensive experiments have been conducted on several publicly available benchmarks and detailed analysis are reported on issues such as the effectiveness of each key components, and the computation cost. These experiments, together with a carefully designed user study, consistently validate the effectiveness and robustness of our algorithm in comparison to the state-of-the-arts.

6 Acknowledges

This work was supported in part by the Beijing Natural Science Foundation under Grant 4182056, the National Basic Research Program of China under grant 2013CB328805, and the Fok Ying-Tong Education Foundation for Young Teachers. Specialized Fund for Joint Building Program of Beijing Municipal Education Commission. Ling was supported in part by US NSF Grants 1618398, 1449860 and 1350521.

References

[1] L. Marchesotti, C. Cifarelli, and G. Csurka, "A framework for visual saliency detection with applications to image thumbnailing," in *Proceedings of the IEEE International Conference on Computer Vision*, 2009, pp. 2232–2239.

[2] J. Sun and H. Ling, "Scale and object aware image thumbnailing," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 135–153, 2013.

[3] J. Chen, G. Bai, S. Liang, and Z. Li, "Automatic image cropping : A computational complexity study," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 507–515.

[4] M. Nishiyama, T. Okabe, Y. Sato, and I. Sato, "Sensation-based photo cropping," in *Proceedings of the ACM International Conference on Multimedia*, 2009, pp. 669–672.

[5] J. Yan, S. Lin, S. Bing Kang, and X. Tang, "Learning the change for automatic image cropping," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 971–978.

[6] C. Fang, Z. Lin, R. Mech, and X. Shen, "Automatic image cropping using visual composition, boundary simplicity and content preservation models," in *Proceedings of the ACM International Conference on Multimedia*, 2014, pp. 1105–1108.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *European Conference on Computer Vision*, 2014, pp. 346–361.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[9] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[10] Y.-L. Chen, T.-W. Huang, K.-H. Chang, Y.-C. Tsai, H.-T. Chen, and B.-Y. Chen, "Quantitative analysis of automatic image cropping algorithms: A dataset and comparative study," in *IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2017, pp. 226–234.

[11] D. Gao and N. Vasconcelos, "Discriminant saliency for visual recognition from cluttered scenes," in *Advances in Neural Information Processing Systems*, 2005, pp. 481–488.

[12] A. K. Mishra, Y. Aloimonos, L. F. Cheong, and A. Kassim, "Active visual segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 639–653, 2012.

[13] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3395–3402.

[14] W. Wang, J. Shen, R. Yang, and F. Porikli, "Saliency-aware video object segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 1, pp. 20–33, 2018.

[15] W. Wang, J. Shen, H. Sun, and L. Shao, "Video co-saliency guided co-segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[16] B. Suh, H. Ling, B. B. Bederson, and D. W. Jacobs, "Automatic thumbnail cropping and its effectiveness," in *Proceedings of the Annual ACM Symposium on User Interface Software and Technology*, 2003, pp. 95–104.

[17] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[18] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Advances in Neural Information Processing Systems*, 2006, pp. 155–162.

[19] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems*, 2007, pp. 545–552.

[20] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, no. 7, pp. 32–32, 2008.

[21] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Advances in Neural Information Processing Systems*, 2009, pp. 681–688.

[22] R. Cong, J. Lei, H. Fu, Q. Huang, X. Cao and C. Hou, "Co-Saliency Detection for RGBD Images Based on Multi-Constraint Feature Matching and Cross Label Propagation," *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 568-579, 2018.

[23] D. Gao, V. Mahadevan, and N. Vasconcelos, "The discriminant center-surround hypothesis for bottom-up saliency," in *Advances in Neural Information Processing Systems*, 2008, pp. 497–504.

[24] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proceedings of the IEEE International Conference on Computer Vision*, 2009, pp. 2106–2113.

[25] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 438–445.

[26] M. Carrasco, "Visual attention: The past 25 years," *Vision Research*, vol. 51, no. 13, pp. 1484–1525, 2011.

[27] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185–207, 2013.

[28] W. Wang and J. Shen, "Deep Visual Attention Prediction," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2368-2378, 2018.

[29] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2798–2805.

[30] W. Wang, J. Shen, F. Guo, M.-M. Cheng, and A. Borji, "Revisiting video saliency: a large-scale benchmark and a new model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[31] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu, "Predicting eye fixations using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 362–370.

[32] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "SALICON: Saliency in context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1072–1080.

[33] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor, "Shallow and deep convolutional networks for saliency prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 598–606.

[34] W. Wang, J. Shen, and L. Shao, "Consistent video saliency using local gradient flow optimization and global refinement," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4185–4196, 2015.

[35] T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[36] W. Wang, J. Shen, and A. Borji, "Salient object detection driven by fixation prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[37] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1597–1604.

[38] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 280–287.

[39] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 38–49, 2018.

[40] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5706–5722, 2015.
[41] Y. Deng, C. C. Loy, and X. Tang, "Image aesthetic assessment: An experimental survey," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 80–106, 2017.
[42] 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*, 2006, pp. 288–301.
[43] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 419–426.
[44] S. Dhar, V. Ordonez, and T. L. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1657–1664.
[45] M. Nishiyama, T. Okabe, I. Sato, and Y. Sato, "Aesthetic quality classification of photographs based on color harmony," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 33–40.
[46] W. Luo, X. Wang, and X. Tang, "Content-based photo quality assessment," in *Proceedings of the IEEE International Conference on Computer Vision*, 2011, pp. 2206–2213.
[47] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *Proceedings of the IEEE International Conference on Computer Vision*, 2011, pp. 1784–1791.
[48] H.-H. Su, T.-W. Chen, C.-C. Kao, W. H. Hsu, and S.-Y. Chien, "Scenic photo quality assessment with bag of aesthetics-preserving features," in *Proceedings of the ACM International Conference on Multimedia*, 2011, pp. 1213–1216.
[49] 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*, 2014, pp. 457–466.
[50] H. Tang, N. Joshi, and A. Kapoor, "Blind image quality assessment using semi-supervised rectifier networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2877–2884.
[51] 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 IEEE International Conference on Computer Vision*, 2015, pp. 990–998.
[52] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *European Conference on Computer Vision*, 2016, pp. 662–679.
[53] L. Mai, H. Jin, and F. Liu, "Composition-preserving deep photo aesthetics assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 497–506.
[54] W. Wang, J. Shen, Y. Yu, and K.-L. Ma, "Stereoscopic thumbnail creation via efficient stereo saliency detection," *IEEE Transactions on Visualization and Computer Graphics*, 2016.
[55] B. Cheng, B. Ni, S. Yan, and Q. Tian, "Learning to photograph," in *Proceedings of the ACM International Conference on Multimedia*, 2010, pp. 291–300.
[56] L. Zhang, M. Song, Q. Zhao, X. Liu, J. Bu, and C. Chen, "Probabilistic graphlet transfer for photo cropping," *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 802–815, 2013.
[57] W. Wang and J. Shen, "Deep cropping via attention box prediction and aesthetics assessment," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.
[58] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.
[59] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.
[60] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *European Conference on Computer Vision*, 2016, pp. 21–37.
[61] N. Murray, L. Marchesotti, and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2408–2415.
[62] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
[63] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.
[64] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*, 2014, pp. 740–755.
[65] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.
[66] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 194–201, 2012.
[67] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 1915–1926, 2012.
[68] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rating image aesthetics using deep learning," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2021–2034, 2015.
[69] J. Park, J.-Y. Lee, Y.-W. Tai, and I. S. Kweon, "Modeling photo composition and its application to photo re-arrangement," in *Proceedings of the IEEE International Conference on Image Processing*, 2012, pp. 2741–2744.
[70] J. Zhang and S. Sclaroff, "Exploiting surroundedness for saliency detection: a boolean map approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 5, pp. 889–902, 2016.
[71] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *International Conference on Machine Learning*, 2014, pp. 647–655.

Wenguan Wang received the B.S. degree in computer science and technology from the Beijing Institute of Technology in 2013. He is currently working toward the Ph.D. degree in the School of Computer Science, Beijing Institute of Technology, Beijing, China. His current research interests include computer vision and deep learning. He received the Baidu Scholarship in 2016.

Jianbing Shen (M'11-SM'12) is a Professor with the School of Computer Science, Beijing Institute of Technology. He has published about 100 journal and conference papers such as *IEEE TPAMI*, *IEEE CVPR*, and *IEEE ICCV*. He has obtained many flagship honors including the Fok Ying Tung Education Foundation from Ministry of Education, the Program for Beijing Excellent Youth Talents from Beijing Municipal Education Commission, and the Program for New Century Excellent Talents from Ministry of Education. His research interests include computer vision and deep learning.

Haibin Ling received B.S. and MS degrees from Peking University, China, in 1997 and 2000, respectively, and the PhD degree from the University of Maryland in 2006. From 2000 to 2001, he was an assistant researcher at Microsoft Research Asia. From 2006 to 2007, he worked as a postdoctoral scientist at the University of California Los Angeles. After that, he joined Siemens Corporate Research as a research scientist. Since 2008, he has been with Temple University where he is now an Associate Professor. He received the Best Student Paper Award at the ACM UIST in 2003, and the NSF CAREER Award in 2014. He is an Associate Editor of *IEEE Trans. on PAMI*, *Pattern Recognition*, and *CVIU*, and served as Area Chairs for *CVPR* 2014 and *CVPR* 2016.