Deep Cropping via Attention Box Prediction and Aesthetics Assessment

Wenguan Wang, and Jianbing Shen*
Beijing Lab of Intelligent Information Technology,
School of Computer Science, Beijing Institute of Technology, China

Abstract

We model the photo cropping problem as a cascade of attention box regression and aesthetic quality classification, based on deep learning. A neural network is designed that has two branches for predicting attention bounding box and analyzing aesthetics, respectively. The predicted attention box is treated as an initial crop window where a set of cropping candidates are generated around it, without missing important information. Then, aesthetics assessment is employed to select the final crop as the one with the best aesthetic quality. With our network, cropping candidates share features within full-image convolutional feature maps, thus avoiding repeated feature computation and leading to higher computation efficiency. Via leveraging rich data for attention prediction and aesthetics assessment, the proposed method produces high-quality cropping results, even with the limited availability of training data for photo cropping. The experimental results demonstrate the competitive results and fast processing speed (5 fps with all steps).

1. Introduction

Consider Fig. 1 (a). How can we determine an appropriate crop for this picture? It seems to be a natural choice that people first define a crop that covers the desired or important region, and then, iteratively adjust the position, size and ratio of the initial crop window until achieving visual-quality-inspired result. This *determining-adjusting* cropping strategy brings two advantages: (1) considering both attention and aesthetics in a cascaded way; and (2) high computation efficiency since the searching space of the best crop is only limited to the surrounding of the initial crop area. Interestingly, however, most previous cropping approaches are
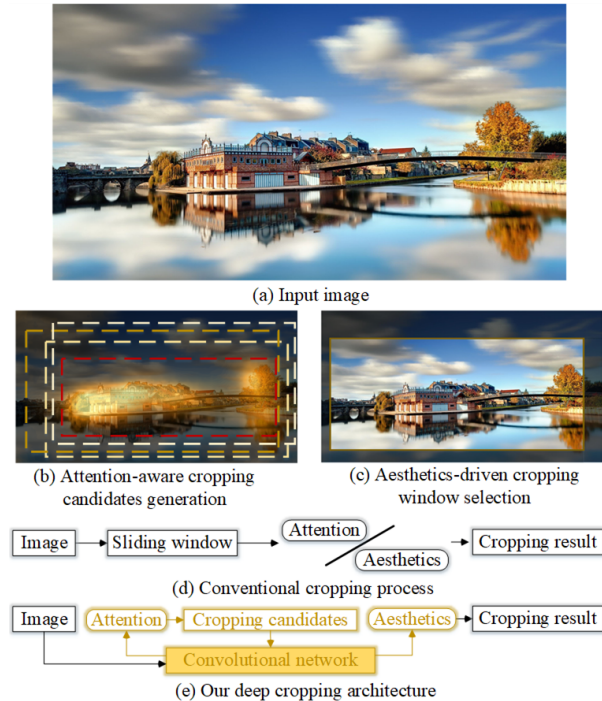
Figure 1: (a)-(c) Flowchart of our method. (d) Conventional methods apply sliding-judging cropping strategy, which is time-consuming and violates natural cropping procedure. (e) Our method works as a cascade of attention-aware crop candidates generation and aesthetics-based crop window selection, which handles photo cropping in a more natural manner and is achieved by a neural network.

proceeded in another way. They usually generate a large number of sliding windows with various ratios and sizes over all the positions, and find the optimal subview via repeatedly computing attention scores [29, 40, 47, 3], or analyzing aesthetics [32, 48] for all the sliding windows. This *sliding-judging* strategy, as depicted in Fig. 1 (d), is companied with heavy computation load, since the searching space would span all the possible subviews of the whole image. Besides, compared with repeatedly calculating attention and/or aesthetics scores over all the crop windows,

arranging these two items in a sequential order would be a more reasonable and time-saving choice.

In this paper, we design a deep learning based cropping method, which models the cropping tasks as attention bounding box *regression* and aesthetics *classification* problems. The network is learned for directly determining the attention box that covers visually important area (the red rectangle in Fig. 1 (b)), which seems like people first placing a crop to cover important region. Then the method generates cropping candidates (the yellow rectangles in Fig. 1 (b)) around the attention box and selects the one with the highest aesthetics value as final crop (Fig. 1 (c)), as the process of human iteratively adjusting initial crop and selecting the most beautiful crop window.

The proposed method approaches cropping task in a more natural and efficient way, which has the following major characteristics and contributions:

Natural and unified deep cropping scheme. The cropping procedure is arranged as a determining-adjusting process, where attention-guided cropping candidates generation is cascaded by aesthetics-aware crop window selection, as demonstrated in Fig. 1 (e). The tasks of attention box predication and aesthetics assessment are achieved in a deep learning model, where attention information is exploited for avoiding discarding important information, while the aesthetics assessment is employed for ensuring the high aesthetic value of cropping results. The deep learning model is based on *fully convolutional neural network*, which naturally supports input images of arbitrary sizes, thus avoiding undesired deformation for evaluating aesthetic quality.

High computation efficiency. Three strategies for enhancing computational efficiency are proposed to achieve a fast processing speed of 5 fps. First, instead of searching all the possible positions in an image domain via sliding window, the approach directly regresses the attention box and generates far less number of cropping candidates ($\sim$1000) around the visual important areas. Second, the sub-networks of attention box prediction and aesthetics assessment share several convolutional layers in the bottom. The marginal cost for computing aesthetics estimate is decreased via sharing convolutions with attention prediction task at test-time. Third, the approach inherits the spirit of recent object detection algorithms [13, 35, 9], which is trained to share convolutional features among cropping candidates on the feature maps. The convolutional layers are only performed once on the entire image (regardless of the number of cropping candidates), and then convolutional features of cropping candidates are extracted from feature maps, which avoiding applying the network to each cropping candidate for repeatedly computing features.

Learning without sufficient cropping annotation. For applying deep leaning for photo cropping, an important practical catch to that solution is training data availability. The datasets for photo cropping are small-scale in deep learning terms, and primarily support evaluation. Besides, the photo cropping sometimes is a quite subjective problem which is difficult to offer a clear answer for what is a 'groundtruth' crop. While the groundtruth for photo cropping is difficult to access, datasets for human gaze prediction and photo aesthetics assessment are more easily to obtain. In our method, the cropping task is explicitly achieved via learning neural network on existing rich and high-quality data for visual attention prediction and aesthetics assessment.

2. Related Work

In this section, we give a brief overview of recent works in three lines: visual attention prediction, aesthetics assessment and photo cropping.

2.1. Visual Attention Prediction

Visual attention prediction aims to predict scene locations where a human observer may fixate. Early attention models [16, 2] are typically based on various low-level features (*e.g.*, color, intensity, orientation), operating and combining them at multiple scales to form a saliency map. In addition to low-level features, some approaches [19, 1] try to employ high-level features from person or face detectors learned from specific computer vision tasks. Recently, driven by the success of deep learning in object recognition, many deep learning based attention models [42, 23, 18, 33] are proposed, and generally give impressive results. The output of traditional attention methods is usually a grayscale image that represents the visual importance of each corresponding pixel in the image. However, in our approach, we try to predict an attention bounding box, which covers the most informative regions of the image.

2.2. Aesthetics Assessment

The main goal of aesthetics assessment is to imitate human interpretation of the beauty of natural images. Many methods have been proposed for this topic, we refer the reader to [5] for a more detailed survey. Traditionally, aesthetic quality analysis is viewed as a binary classification problem of predicting high- and low- quality images. Extracting visual features and then employing various machine learning algorithms to predict photo aesthetic values is a common pipeline in this research area.

Early methods [4, 20, 6] manually designed aesthetics features according to photographic rules or practices, such as the rule of thirds and visual balance. Instead of using hand-crafted features, other approaches [30, 38] have been developed to leverage more generic image descriptors, such as Fisher Vector and bag of visual words, which are previously used for image classification but also capable of capturing aesthetic properties. In more recent work

[25, 41, 27, 21, 28], deep learning methods have been used to aesthetics assessment and have shown promising results.

2.3. Photo Cropping

Cropping is an important operation for improving visual quality of digital photos, which cuts away unwanted areas outside of a selected rectangular region. A lot of methods have been proposed towards automating this task. These methods, in general, can be categorized into *attention-based* or *aesthetics-based* approaches. The attention-based approaches [29, 40, 3] focus on preserving the main subject or visually important area in the scene after cropping. These methods usually place the crop window over the most visually significant regions according to certain attention scores [43, 44, 45, 46]. The other major direction of cropping methods is aesthetics-based approach that emphasizes the general attractiveness of the cropped image. Those aesthetics-based approaches [32, 48] are centered on composition-related image properties. Taking various aesthetical factors into account, they try to find the cropping candidate with the highest quality score.

In this paper, we consider both attention and aesthetics information, which are arranged in a natural and cascaded manner. The proposed method approaches photo cropping as a cascade of generating cropping candidates via attention box prediction and selecting best crop according to aesthetics criteria. Our method shares the spirit of recent object detection algorithms [13, 35, 9], one branch of our network learns to predict the bounding box covers visually important area, while the other one tries to analyze aesthetic value.

3. Our Approach

The cropping algorithm is decomposed into two cascaded stages, namely, attention-aware cropping candidates generation (Sec. 3.1) and aesthetics-based crop window selection (Sec. 3.2). It infers initial crop as a bounding box covering the most visually important area, and then selects the best crop with highest aesthetic quality from a few crop candidates generated around the initial crop. We design a deep learning model that has two sub-networks: Attention Box Prediction (ABP) network and Aesthetics Assessment (AA) network, for achieving two key subtasks in above cropping process: (1) attention box prediction for determining the initial crop; and (2) aesthetics assessment for determining the final crop. Those 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 Sec. 3.3, we will 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
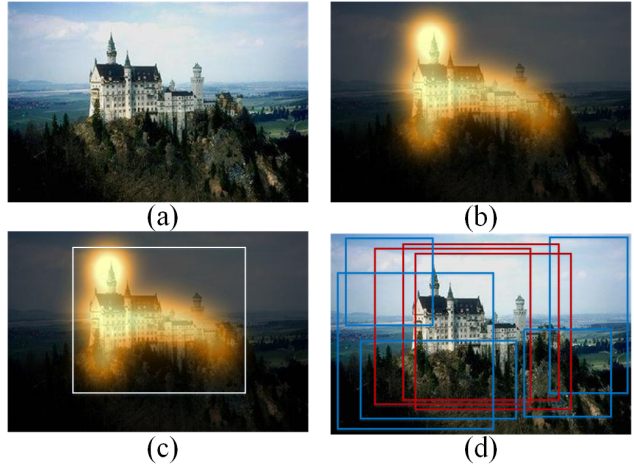

Figure 2: (a) Input image. (b) Attention map. (c) Ground truth attention box generation via [3]. (d) Positive (red) and negative (blue) defaults boxes are generated for training ABP network according to ground truth attention box.

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

The initial crop can be viewed as a rectangle that preserves the most informative part of the image while has minimum area. This optimal rectangle searching problem is a common task for attention-based cropping methods. Let $P \in [0, 1]^{w \times h}$ be an attention mask, we first define a set of crop windows $\mathfrak{W}$:

$$\mathfrak{W} = \{W | \sum_{x \in W} P(x) > \lambda \sum_{x} P(x)\}, \quad (1)$$

where $\lambda \in [0, 1]$ is a fraction threshold. Then the optimum rectangle $\dot{W}$ is defined as:

$$\dot{W} = \operatorname*{argmin}_{W \in \mathfrak{W}} |W|. \quad (2)$$

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

Differently, we design a neural network for directly predicting such attention box. Given a training sample (I, G) consisting of an image I of size $w \times h \times 3$ (Fig. 2 (a)), and a groundtruth attention map $G \in [0, 1]^{w \times h}$ (Fig. 2 (b)), the optimum rectangular $\dot{W}$ defined in Equ. 2 is computed as the groundtruth attention prediction box. Here we apply
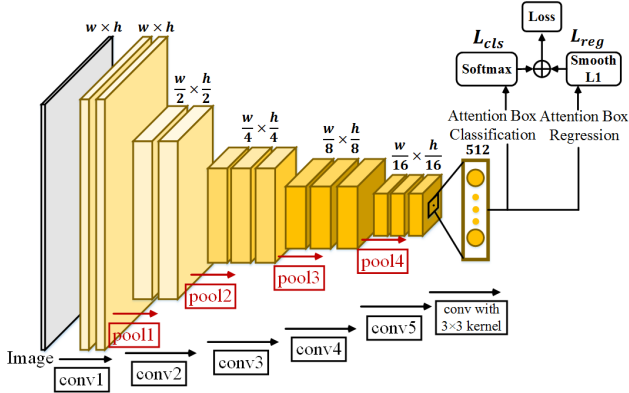
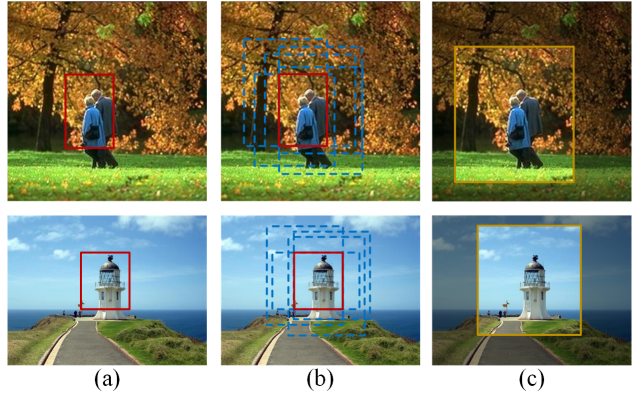Figure 3: Architecture of Attention Box Prediction (ABP) network.

Figure 4: (a) Initial crop (red rectangle) predicted via ABP network. (b) Cropping candidates (blue rectangles) generated around the initial crop. (c) Final crop selected as the candidate with highest aesthetic score from AA network.

[3] for generating $\dot{W}$ over G (Fig. 2 (c)) for computation efficiency. We set $\lambda = 0.9$ for preserving most informative areas. Then the task of attention box prediction can be achieved via bounding box regression as object detection [13, 35, 9]. Note that any other attention scores can also be used for generating groundtruth bounding box for training the ABP network.

Fig. 3 illustrates the architecture of ABP network. The bottom of this network is a stack of convolutional layers, which are borrowed from the first five convolutional blocks of VGGNet [37]. With the last convolutional layer, we slide a small network with 3×3 kernel over its convolutional feature map, thus generating $512-d$ feature for each sliding location. The feature vector is further fed into two fully-connected layers: box-regression layer for predicting attention bounding box; box-classification layer for determining the box whether 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 [35].

During training, we need to determine which default boxes correspond to the groundtruth attention box and train the network accordingly. We assign the default box which has the highest Intersection-over-Union (IoU) with the groundtruth box or with IoU higher than 0.7 as a positive label ($c = 1$). We assign the default box that has a IoU lower than 0.3 a negative label ($c = 0$) and drop other default boxes. The above process is illustrated in Fig. 2 (d). For the preserved boxes, we define $\bar{p}_i^c \in \{1, 0\}$ as an indicator for the label of i-th box and vector $\bar{t}$ as a four-parameterized 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 i-th default box. With above definition, the ABP network is trained via minimizing the following loss function derived from object detection [10, 35, 24]:

$$\mathcal{L}(p, t) = \sum_i \mathcal{L}_{cls}(p_i, \bar{p}_i) + \sum_i \bar{p}_i^1 \mathcal{L}_{reg}(t_i, \bar{t}). \quad (3)$$

The classification loss $\mathcal{L}_{cls}$ is the softmax loss over confidences of two classes (attention box or not). The regression loss $\mathcal{L}_{reg}$ is a Smooth L1 loss [10], between the predicated box and the ground truth attention box, which is only activated for positive default boxes.

With the ABP network trained on existing attention prediction datasets, it learns to generates reliable attention boxes. Then we select the one with the highest prediction score (p_i^1) as the initial crop. This initial crop covers the most informative part of the image, which likes human placing a crop around the desired area (Fig. 4 (a)). Next, we generate a set of cropping candidates around the initial crop, as the human adjusting the location, size and ratio of the initial crop. A rectangular can be uniquely determined via the coordinates of its top-left and right-bottom corners. For the top-left corner of the initial crop, we define a set of offsets: $\{-40, -32, \cdots, -8, 0\}$ in x- and y-axis. Similarly, a set of offsets: $\{0, 8, ..., 32, 40\}$ in x- and y-axis is also defined for the bottom-right corner. Via adding the top-left and bottom-right corners with corresponding pre-defined offsets [1], we generate $6^4 = 1296$ cropping candidates in total, which is far less than the sliding windows needed for traditional cropping methods. Each of crop candidates is designed for covering the whole initial crop area, since the initial crop is a minimum visually importance-preserved rectangle that should be maintained in cropping process (Fig. 4 (b)).
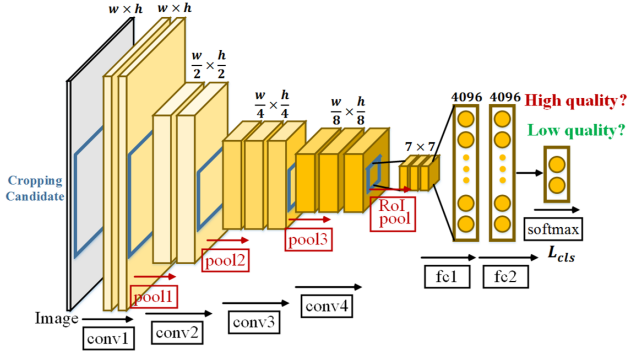
Figure 5: Architecture of Aesthetics Assessment (AA) network.

3.2. Aesthetics-based Crop Window Selection

With our attention-aware cropping candidates by ABP network, we next select the most aesthetics-inspired one as the final crop. It is important to consider aesthetics for photo cropping task, since beyond preserving the important content, a nice crop should also deliver pleasant viewing experience. For analyzing the aesthetic quality of each cropping candidates, one choice is training an aesthetics assessment network, and iteratively applying forward-propagation for each crop candidate over this network when cropping. Obviously, this strategy is straightforward but time-consuming. Inspired by the recent advantages of object detection, which share convolutional features between regions, we build a network that analyzes aesthetic values of all cropping candidates simultaneously.

We achieve this via an Aesthetics Assessment (AA) network (Fig. 5), 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 former four convolutional blocks of VGGNet [37] without $pool4$ layer. Here we adopt a relatively shallow network mainly due to two reasons. First, aesthetics assessment is a relatively easier problem (high quality *vs* low quality) compared with image classification (with 1000 classes). Secondly, for an image with the 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 Sec. 3.1.

Then, on the top of the last convolutional layer, we adopt Region of Interest (RoI) pooling layer [35], which is a special case of spatial pyramid pooling (SPP) layer [13], to extract a fixed-length feature vector from the final convolutional feature map. The RoI pooling layer uses max pooling to convert the features inside any crop candidate into a small feature map with a fixed-dimensional vector, which is further fed into a sequence of fully-connected layer for

---

¹Since we resize the input image with $min(w, h) = 224$, we find the largest offset (40) is enough.

aesthetic quality classification. This operation allows us to operate image with arbitrary aspect ratios, thus avoiding undesired deformation in aesthetics assessment. With a crop candidate with size of $w' \times h'$, RoI pooling layer divides it into $n \times n$ spatial bins and applies max-pooling for the features within each bins. Here we set $n = 7$.

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

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

$$\mathcal{L}_{cls}(q, \bar{q}) = -\frac{1}{N} \sum_i \sum_{c \in \{\mathbf{1}, \mathbf{0}\}} \bar{q}_i^c log(\widehat{q}_i^c),$$

$$where \ \widehat{q}_i^c = exp(q_i^c) \Big/ (exp(q_i^{\mathbf{1}}) + exp(q_i^{\mathbf{0}})). \tag{4}$$

With the cropping candidates generated from APB network, the AA network is capable of producing their aesthetics-quality scores ($\{q_i^{\mathbf{1}}\}$), where the one with the highest score is selected as the final crop (Fig. 4 (c)).

3.3. Implementation Details

Training Two large-scale datasets: SALICON [18] and AVA [31], are used for training our model. SALICON is used for training our ABP network. It contains 15000 natural images with eye fixation annotations which are simulated through mouse movements of users on blurred images. For obtaining groundtruth attention box, we follow the instructions of [18] for transferring the binary mouse-clicking map into grey-scale human attention map, and then we apply [3] for generating attention bounding box according to Equ. 2 with $\lambda = 0.9$. The AVA dataset is the largest publicly available aesthetics assessment benchmark, which provides 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 by [25, 27, 28, 31], about 230,000 images are used for training our AA network. More specially, images with mean ratings smaller than 5 are assigned as low quality and those with mean ratings larger than or equal to 5 are labeled as high quality.

Our two sub-networks are trained simultaneously. In each training iteration, we use a min-batch of 4 images, 2 of which are from SALICON dataset with the groundtruth attention boxes and the rest from AVA dataset with aesthetics quality groundtruth. Before feeding the input images and

(a) Images with highest aesthetics values   (b) Images with lowest aesthetics values
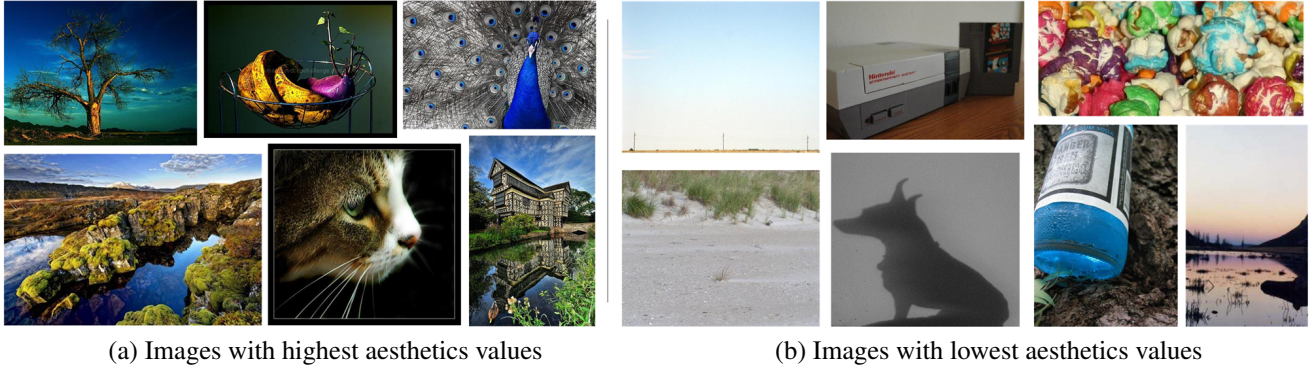
Figure 6: Aesthetics assessment results via our AA network. The test images with the highest predicted aesthetics values and those with the lowest predicted aesthetics values are presented in (a) and (b), respectively.

ground-truth to the network, we scale the images such that the smaller dimension is 224. Since the bottom two convolutional blocks ($conv1$ and $conv2$) are shared between both the tasks of attention box prediction and esthetics assessment, they are trained for the two tasks simultaneously using all the images in the batch. For the layers specialized to each of the sub-networks are trained using only those images in the batch having the corresponding ground-truth.

Both ABP and AA networks are initialized from the weights of VGGNet [37], which is pre-trained on large-scale image classification dataset [36]. Our model is implemented with the popular Caffe library [17] and trained with stochastic gradient descent. The networks were trained over 200K iterations where we use momentum of 0.9 and weight decay of 0.0001, which is reduced by a factor of 0.1 at every 10K iterations.

Testing For training, our two sub-networks are trained in parallel strategy, while for testing, they work in a cascaded way. With a given image (resized with $min(w, h) = 224$) for cropping, we first gain a set of attention boxes generated via forward propagation on APB network. Then the initial crop was selected as the one with the highest score of attention box prediction. After that, a set of cropping candidates are generated around the initial crop. Since the two convolutional blocks at the bottom are shared between ABP and AA networks, we directly feed the cropping candidates and the convolutional feature of last layer of $conv2$ into AA network. Finally, the final crop is selected as the cropping candidate with best aesthetic quality. The cropping model achieves a fast speed of 5 fps.

4. Experimental results

In this section, we first examine the performance of our ABP and AA networks on their specific tasks. The goal of these experiments is to investigate the effectiveness of individual components instead of comparing them with the state-of-the-art. Then, we evaluate the performance of our

whole cropping model on two widely used photo cropping datasets with other competitors.

4.1. Evaluation for ABP and AA Networks

Performance of ABP Network We first evaluate the performance of ABP network on PASCAL dataset [22], which is widely used for attention prediction. This dataset contains totally 850 natural images from PASCAL 2010 [7], with the eye fixations during 2 seconds of 8 different subjects. With the binary eye fixation images, we follow [22] to generate gray-scale attention map. Then, as the way described in Sec. 3.3, we generate groundtruth attention box for each image. We consider eight state-of-the-art attention models: ITTI [16], AIM [2], GBVS [12], SUN [49], DVA [15], SIG [14], CAS [11] and SalNet [33]. Then we extract the attention boxes of above methods via the same strategy used for generating groundtruth bounding box. We opt for the Intersection over Union (IoU) score for quantifying the quality of extracted attention boxes. The quantitative results are illustrated in Table 1. As seen, our attention box prediction results are more accurate than previous attention models, since our ABP network is specially designed for this task.

Method	**Ours**	ITTI[16]	AIM [2]	GBVS[12]	SUN[49]
IoU	**0.517**	0.318	0.327	0.319	0.273

Method	**Ours**	DVA[15]	SIG[14]	CAS [11]	SalNet [33]
IoU	**0.517**	0.346	0.272	0.356	0.379

Table 1: Attention box prediction with IoU for PASCAL [22].

Performance of AA Network We adopt the testing set of AVA dataset [31], which is mentioned in Sec. 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: AVA [31],

Method	Ours	AVA[31]	RAP-DCNN[25]	RAP-RDCNN[25]
Accuracy	**0.769**	0.667	0.732	0.745

Method	Ours	RAP2[26]	DMA-SPP[27]	DMA[27]
Accuracy	**0.769**	0.754	0.728	0.745

Method	Ours	DMA-Alex[27]	ARC[21]	CPD[28]
Accuracy	**0.769**	0.754	0.773	0.774

Table 2: Aesthetics assessment accuracy for AVA [31].

RAP [25], RAP2 [26], DMA [27], ARC [21] and CPD [28], where AVA is based on manually designed features while other methods are based on deep learning model. As shown in Table 2, our AA network is struggle to achieve state-of-the-art performance due to relatively simple network architecture. In Fig. 6, we present some examples of the test images that are considered of the highest and lowest aesthetics values by our AA network.

Conclusion Overall, our two sub-networks generate the promising results or compete with existing top-performance approaches. Considering the shared convolutional layers in the bottom of these two networks, our model achieves a good tradeoff between performance and computation efficiency. More important, the robustness of those two basic components greatly contributes the high-quality of our crop suggestions, which will be detailed in next section.

4.2. Evaluation for Photo Cropping

We evaluate our whole cropping model on two public image cropping datasets, including Image Cropping Dataset from MSR (MSR-ICD) [48] and FLMS [8]. The MSR-ICD dataset includes 950 images and each image is carefully cropped by 3 experts. The FLMS dataset contains 500 natural images which are 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.

We adopt the same evaluation metrics as [48], $i.e.$, IoU score and Boundary Displacement Error (BDE), to measure the cropping accuracy of image croppers. BDE is defined as the mean normalized displacement of four edges between the cropping box and the groundtruth rectangles.

Method	Photographer 1		Photographer 2		Photographer 3	
	IoU↑	BDE↓	IoU↑	BDE↓	IoU↑	BDE↓
ATC [39]	0.605	0.108	0.628	0.100	0.641	0.095
AIC [3]	0.469	0.142	0.494	0.131	0.512	0.123
LCC [48]	0.748	0.066	0.728	0.072	0.732	0.071
MPC [34]	0.603	0.106	0.582	0.112	0.608	0.110
SPC [32]	0.396	0.177	0.394	0.178	0.385	0.182
ARC [21]	0.448	0.163	0.437	0.168	0.440	0.165
Ours	**0.813**	**0.030**	**0.806**	**0.032**	**0.816**	**0.032**

Table 3: Cropping results with IoU and BDE on MSR-ICD [48].

We compare our cropping method with two main categories of image cropping methods, $i.e.$, *attention-based* and *aesthetics-based* methods. For attention-based method, we select ATC [39] which is a classical image thumbnail cropping method. We also use AIC as a baseline, which is obtained via equipping crop window researching method [3] with top-performing saliency detection method. We

Method	Ours	ATC [39]	AIC [3]	LCC [48]	MPC [34]	VBC [8]
IoU↑	**0.81**	0.72	0.64	0.63	0.41	0.74
BDE↓	**0.057**	0.063	0.075	-	-	-

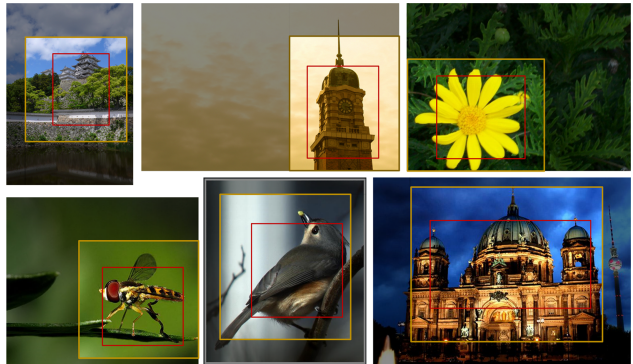Table 4: Cropping results with IoU and BDE on FLMS [8].

Figure 7: Qualitative results on MSR-ICD [48] and FLMS [8] datasets. The red rectangles indicate the initial crop generated via ABP network, and the yellow windows correspond to the final crop selected via AA network.

apply context-aware saliency [11] and optimal parameters, as suggested by [3], for maximizing its performance. For aesthetics-based method, we select LCC [48], MPC [34], and VBC [8]. We also consider SPC, which is an advanced version of [32], as described in [48]. Additionally, we adopt a recent aesthetics ranking method [21] combined with sliding window strategy as a baseline: ARC. We select the crop as the one with the highest ranking score from sliding windows. The comparison results on MSR-ICD and FLMS datasets are demonstrated in Table 3 and Table 4, respectively. As seen, our cropping method achieves the best performance in both datasets. Qualitative results on MSR-ICD and FLMS datasets are presented in Fig. 7.

5. Conclusions

In this work, we propose a deep learning based photo cropping approach, driven by human attention box prediction and aesthetics assessment. The proposed deep model is decomposed into two sub-networks: Attention Box Prediction (ABP) network and Aesthetics Assessment (AA) network, which share multiple convolution layers at the bottom. The proposed method approaches photo cropping in a determining-adjusting manner. It infers initial

crop as a bounding box covering the visually important area (attention-aware determining), and then selects the best crop with highest aesthetic quality from a few cropping candidates generated around the initial crop (aesthetic-based adjusting). Our extensive experimental analyses demonstrate that our solution achieves superior performance in comparison to the state-of-the-art.

References

[1] A. Borji. Boosting bottom-up and top-down visual features for saliency estimation. In *CVPR*, 2012.

[2] N. Bruce and J. Tsotsos. Saliency based on information maximization. *NIPS*, 2006.

[3] J. Chen, G. Bai, S. Liang, and Z. Li. Automatic image cropping : A computational complexity study. In *CVPR*, 2016.

[4] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *ECCV*, 2006.

[5] Y. Deng, C. C. Loy, and X. Tang. Image aesthetic assessment: An experimental survey. *arXiv preprint arXiv:1610.00838*, 2016.

[6] S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *CVPR*, 2011.

[7] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes (VOC) challenge. *IJCV*, 2010.

[8] C. Fang, Z. Lin, R. Mech, and X. Shen. Automatic image cropping using visual composition, boundary simplicity and content preservation models. In *ACMMM*, 2014.

[9] R. Girshick. Fast R-CNN. In *ICCV*, 2015.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[11] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *IEEE PAMI*, 2012.

[12] J. Harel, C. Koch, P. Perona, et al. Graph-based visual saliency. In *NIPS*, 2006.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*, 2014.

[14] X. Hou, J. Harel, and C. Koch. Image signature: Highlighting sparse salient regions. *IEEE PAMI*, 2012.

[15] X. Hou and L. Zhang. Dynamic visual attention: Searching for coding length increments. In *NIPS*, 2009.

[16] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE PAMI*, 1998.

[17] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[18] M. Jiang, S. Huang, J. Duan, and Q. Zhao. SALICON: Saliency in context. In *CVPR*, 2015.

[19] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV*, 2009.

[20] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *CVPR*, 2006.

[21] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*, 2016.

[22] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille. The secrets of salient object segmentation. In *CVPR*, 2014.

[23] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu. Predicting eye fixations using convolutional neural networks. In *CVPR*, 2015.

[24] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. In *ECCV*, 2016.

[25] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang. RAPID: Rating pictorial aesthetics using deep learning. In *ACMMM*, 2014.

[26] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang. Rating image aesthetics using deep learning. In *IEEE TMM*, 2015.

[27] 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 *ICCV*, 2015.

[28] L. Mai, H. Jin, and F. Liu. Composition-preserving deep photo aesthetics assessment. In *CVPR*, 2016.

[29] L. Marchesotti, C. Cifarelli, and G. Csurka. A framework for visual saliency detection with applications to image thumbnailing. In *ICCV*, 2009.

[30] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. In *ICCV*, 2011.

[31] N. Murray, L. Marchesotti, and F. Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *CVPR*, 2012.

[32] M. Nishiyama, T. Okabe, Y. Sato, and I. Sato. Sensation-based photo cropping. In *ACMMM*, 2009.

[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 *CVPR*, 2016.

[34] J. Park, J.-Y. Lee, Y.-W. Tai, and I. S. Kweon. Modeling photo composition and its application to photo rearrangement. In *ICIP*, 2012.

[35] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[36] 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. *IJCV*, 2015.

[37] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[38] 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 *ACMMM*, 2011.

[39] B. Suh, H. Ling, B. B. Bederson, and D. W. Jacobs. Automatic thumbnail cropping and its effectiveness. In *ACM UIST*, 2003.

[40] J. Sun and H. Ling. Scale and object aware image thumbnailing. *IJCV*, 2013.

[41] H. Tang, N. Joshi, and A. Kapoor. Blind image quality assessment using semi-supervised rectifier networks. In *CVPR*, 2014.

[42] E. Vig, M. Dorr, and D. Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *CVPR*, 2014.

[43] W. Wang, J. Shen, and F. Porikli. Saliency-aware geodesic video object segmentation. In *CVPR*, 2015.

[44] W. Wang, J. Shen, and L. Shao. Consistent video saliency using local gradient flow optimization and global refinement. *IEEE TIP*, 2015.

[45] W. Wang, J. Shen, L. Shao, and F. Porikli. Correspondence driven saliency transfer. *IEEE TIP*, 2016.

[46] W. Wang, J. Shen, R. Yang, and F. Porikli. Saliency-aware video object segmentation. *IEEE PAMI*, 2017.

[47] W. Wang, J. Shen, Y. Yu, and K.-L. Ma. Stereoscopic thumbnail creation via efficient stereo saliency detection. *IEEE TVCG*, 2016.

[48] J. Yan, S. Lin, S. Bing Kang, and X. Tang. Learning the change for automatic image cropping. In *CVPR*, 2013.

[49] 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*, 2008.