Learning the Relation between Interested Objects and Aesthetic Region for Image Cropping

Peng Lu, Hao Zhang, Xujun Peng, and Xiaofu Jin

***Abstract*—As one of the fundamental techniques for image editing, image cropping discards irrelevant contents and remains the pleasing portions of the image to enhance the overall composition and achieve better visual/aesthetic perception. In this paper, we primarily focus on improving the efficiency of automatic image cropping, and on further exploring its potential in public datasets with high accuracy. From this perspective, we propose a deep learning based framework to learn the objects composition from photos with high aesthetic qualities, where an interested object region is detected through a convolutional neural network (CNN) based on the saliency map. The features of the detected interested objects are then fed into a regression network to obtain the final cropping result. Unlike the conventional methods that multiple candidates are proposed and evaluated iteratively, only a single interested object region is produced in our model, which is mapped to the final output directly. Thus, low computational resources are required for the proposed approach. The experimental results on the public datasets show that as a weakly supervised method, the proposed network outperforms the other weakly supervised methods on FLMS and FCD datasets and achieves comparable results to the existing methods on CUHK dataset. Furthermore, the proposed method is more efficient than these methods, where the processing speed is as fast as 20ms per image.**

***Index Terms*—Deep learning, aesthetics, image composition, convolutional network.**

I. INTRODUCTION

Image cropping, which aims at removing unexpected regions and non-informative noises from a photo/image, by modifying its aspect ratio or through improving the composition, is one of the basic image manipulation processes for graphic design, photography and image editing. Nowadays, with the proliferation of hand-held smart devices, users are more eager to capture photos obtaining not only the theme that the image needs to express but also the good composition and appealing colors, to facilitate semantic searching and to make audiences enjoyable. This trend attracts increasing interests of image cropping from both research community and industries.

However, cropping an image to obtain appropriate composition for achieving better visual quality is notoriously difficult, primarily driven by three facts: (1) to determine the main object/theme of a given image is a nontrivial task, which needs deep domain knowledge and sophisticate skills; (2) assessment of aesthetic of the cropped image is highly subjective such that different viewers might have various opinions for the same cropped photo, or even the same viewer might have opposite feelings for the same image at different time; (3) vast amount of cropping candidate areas can be extracted from the image which causes the solution space is exponentially increased.

P. Lu, H. Zhang and X. Jin are with the School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China. E-mail: plu@bupt.edu.cn

X. Peng is with the Information Sciences Institute, University of Southern California, Marina del Rey, CA, USA. E-mail: xpeng@isi.edu

To tackle these problems, many researchers seek to propose novel approaches to automatically crop images with high aesthetic score. These existing researches can be roughly grouped into four main categories: *sliding-judging* based, *determining-adjusting* based [1], *sequential decision-making* based [2] and *detecting-determining* based methods.

The sliding-judging based approaches normally exhaustively scan the entire image using windows with different size and aspect ratio to produce abundant candidate regions [3], [4]. For each candidate, a classifier or ranker is applied to evaluate its visual/aesthetic quality and the one with the highest score is considered as the optimal cropping result. However, the low computational efficiency of these approaches limits their success. In order to avoid greedily searching against all possible sub-windows, determining-adjusting based approaches attempt to propose a small set of candidate windows with high probabilities to narrow down the searching space for the optimal cropping rectangle. Normally, a seed candidate is initially determined by face, salient object, or attention detection algorithms [5], [6], [7], from which the surrounding areas are scanned and evaluated by classifiers or rankers to select the region with the highest visual/aesthetic quality. Although determining-adjusting based approaches have higher efficiency than sliding-judging based methods, they still encounter the same problems of multiple candidates generation and selection. To avoid evaluating a large amount of proposals, sequential decision-making based approaches use aesthetics aware reward function to guide the searching for cropping windows and decision-making iterations are reduced to as low as dozens for crops prediction [2]. Unlike all existing cropping approaches, by discovering the relation between interested objects and the aesthetic quality of cropped image, detecting-determining based approaches find the optimal cropping rectangle based on the detected interested object region directly without any multiple proposals and evaluations [8].

As can be seen from [8] that by employing interested objects which represent those areas attracting most attentions from the viewers, and its relation to aesthetic areas, the detecting-determining based image cropping approaches demonstrate the promising results for both accuracy and efficiency. However, the brute force searching technique [6] used for the interested object localization (IOL) is still a bottle-neck for the efficiency of this type of methods. And the multiple stages training and inference scheme also limits its applicability. Notably, unlike

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.3029882, IEEE Transactions on Multimedia

the general objects detection approaches, which specially focus on one or multiple objects in its entirety, IOL focuses more on the psychological feelings from the perspective of viewers and is more suitable for the image cropping task. Normally, IOL is computed based on the saliency map detection. Similar to but differ from the salient objects detection [9], [10], [11], saliency map detection attempts to calculate the "saliency map" that simulates the eye movement behaviors of human, but salient objects detection tends to bias to particular objects in the image, which results in different assessment criteria and ground-truth used for these two different tasks. Thus, the saliency map detection generally has higher generalization capability because it is not relied on particular objects. The difference between them is detailed addressed in [1] and [12].

Thus, in this paper we propose a weakly supervised end-to-end image cropping framework to address the problems of detecting-determining based approaches, where the ground truths of cropping bounding boxes are not required in our system. Particularly, the proposed image cropping system uses a deep neural network to extract the saliency map of the image, which is followed by the proposed IOL layer to determine the region containing the interested objects in the image. Then a regression network is employed to map the interested object region to the final cropping rectangle based on its feature. The proposed cropping method only has one pass to achieve the optimal cropping result, without iterative searching or scanning on multiple proposals of different windows, which highly improves the computational efficiency and obtains comparative accuracy performance.

In summary, we make three contributions to the literature:

- We propose a probabilistic framework to model the relationship between the interested objects in the image and the area with high aesthetic quality. Based on this relation, the task of image cropping can be considered as an optimization problem to maximize the joint probability of interested objects and the cropped area with high aesthetic quality for the given image;
- The searching technique which is normally used for the IOL is the main obstacle for the end-to-end training and inference because it is not differentiable. An IOL layer is proposed based on the saliency map to avoid this type of heuristic searching scheme. The proposed layer can effectively find the location of interested objects and is differentiable;
- Based on the proposed probabilistic framework and IOL layer, an end-to-end cropping system is designed, which is not relied on the cropping annotation datasets for training but achieves the state-of-the-art accuracy and high efficiency with 50 frame per second (FPS) on public datasets.

The remainder of this paper is organized as follows. Section II briefly covers the related works to image cropping. Section III is an in-depth introduction of the proposed methodology. The experimental setup, results analysis and discussion are presented in section IV. Finally, we conclude our work in section V.

II. Previous Work

A. Saliency Map Detection

Most vertebrates have the ability to move their eyes and predict fixation with limited time and resources, which enables them to focus on the most informative region and extract the most relevant features for the particular scene [13], [14]. This phenomenon inspired researchers to obtain the cropping areas of image relied on the saliency map prediction.

Generally, saliency map is produced prior to the salient object detection, as demonstrated in [12], where each pixel in the map indicates the confidence of the fixation. In [15], Harel *et al.* proposed a graph-based visual saliency model depended on Markovian chain assumption, which calculated and normalized the activation map by measuring the dissimilarity between neighboring pixels. The reported ROC curve for this method beat the classical attention map detection approach proposed by Itti *et al.*, where multiple empirical features were fed into a neural network to select the proper attended locations [16]. In the same manner, Judd *et al.* defined a set of hand-crafted features to represent low-, mid- and high-level perception of human visual system and fed them into a support vector machine (SVM) to distinguish positive and negative salient pixels [17].

However, the drawback of these mentioned approaches is that strong domain knowledge and experiences are required for design of those hand-tuned features, which is a obstacle to extend their applicability. Therefore, in [18], Vig *et al.* proposed to utilize CNN to learn the representations for salient and non-salient regions. With labeled feature vectors, an L2-regularized, linear, L2-loss SVM was trained to predict the probability of fixation of images in their work.

B. Aesthetic Assessment

Besides salient objects that affect the performance of image cropping, aesthetic, which represents the degree of beauty, is another key factor to determine the quality of cropped images. Early work for aesthetic assessment can be dated back to the researches of color harmony theories [19] and photographic composition [20]. In recent years, many automatic image aesthetic assessment algorithms were proposed, where hand-crafted global features, such as spatial distribution of edges, color distributions, hue count etc. [21] and local features, e.g. wavelet-based texture and shape convexity [22] were employed. To take the advantage of both global and local features, Zhang *et al.* combined structural cues of these two levels for photo aesthetic evaluation [23]. Particularly, graphlet-based local structure descriptors were constructed and projected onto a manifold to preserve the global layout of the image, which was embedded into a probabilistic framework to assess image aesthetic. However, these representations consider whole image indiscriminately ignoring the importance of main subjects in the image. To remedy this problem, Luo *et al.* suggested extracting different subject areas prior to the aesthetic evaluation and treating them using different aesthetic features [24]. Furthermore, genetic image descriptors were also applied to facilitate the aesthetic assessment task. For instance, Marchesotti *et al.* developed two types of local image signatures originated from Bag-of-visual-words and Fisher vectors by incorporating SIFT and color information into them [25].

Under the assumption that semantic recognition task can help the aesthetic assessment, Kao *et al.* proposed a multi-task framework where two tasks were trained simultaneously while the representations were shared by two networks [26]. This idea was also applied by Lu *et al.* in their work of color harmony modeling, which used both bag-of-visual-words features and semantic tag information to boost the aesthetic assessment performance through colors [27]. Moreover, in order to overcome the problem of contaminated tags, a semi-supervised deep active learning algorithm was proposed in [28], where a large set of object patches were extracted and linked to the semantical tags to benefit image aesthetic assessment.

C. Regression Networks

Although photos can be cropped depended on the obtained salient objects only, they are not necessary to be with high aesthetic quality because the aspect of aesthetic is normally ignored for saliency detection. To tackle this problem, one feasible solution is to determine a seed cropping window according to the saliency map and propose a set of candidates around this seed window. Thereafter, every candidate is evaluated by its aesthetic quality and the one with the highest aesthetic score is considered as the final cropping result, as the methods described in [4], [29]. However, iterative assessment for each candidate window's aesthetic score increases the computational complexity. Thus, a more practical and efficient approach is to map the seed salient region to the final cropping window directly using regression models, where the aesthetic information is integrated into the system.

In the field of computer vision [30], [31], regression method is widely used for object detection. Girshick *et al.* combined regions with CNN features (R-CNN) to find different objects in the image [32], where bounding-box regression technique proposed in [33] was applied for a selective search region proposal to refine the detection window . To speed up R-CNN, Girshick improved their work by using a fully connected network to predict the bounding-box regression offsets and confidence of each proposal [34]. Instead of performing classification for detection problem, Redmon *et al.* framed object detection as a regression problem, where the input image was divided into small patches initially and the bounding boxes offsets as well as their probabilities for each class were predicted directly in one neural network, which was called YOLO [35]. The final detections were obtained by merging bounding boxes for the same class. To make YOLO better and faster, Redmon and Farhadi shrink the CNN and used region proposal network (RPN) to generate more anchor boxes for boosting recall and localization accuracy, where regression network was remained for location/confidence prediction [36]. Unlike YOLO, Liu *et al.* detected different objects by evaluating a small set of boxes that were produced through multi-resolution CNN feature maps. The final bounding boxes of objects were also obtained by regressing to offsets for the centers of the default boxes [37]. The similar ideas of using regression networks for objects detection can be found in [38].

D. Image Cropping & Recomposition

As an important procedure to enhance the visual quality of photos, image cropping and recomposition benefit from the development of salient object detection, aesthetic assessment and other computer vision techniques. To combine visual composition, boundary simplicity and content preservation into a photo cropping system, saliency map and salient object were used to encode the spatial configuration and content information, and gradient values were applied to measure the simplicity of image by Fang *et al.* [3]. In this method, image was densely cropped, evaluated and merged by the mentioned features to obtain the optimal cropping results. By segmenting the entire image into small regions, a region adjacency graph (graphlets) was constructed by Zhang *et al.* to represent the aesthetic features of the image, from which the image was cropped through a probabilistic model [39], [40]. Zhang *et al.* also extended the idea of graphlets in the semantic space for image cropping, which was created based on the category information of the images [41]. In the semantic space, semantically representative graphlets were selected sequentially and evaluated by a pre-trained aesthetic prior model to guide the cropping process. Unlike the other algorithms that evaluated multiple candidate cropping areas, Samii *et al.* searched against a high quality image database to find exemplar photos with similar spatial layouts as the query image, and matched the composition of the query image to each of exemplars by minimizing composition distance in a high-level context feature space to calculate the optimal crop areas [42]. In [43], Wang *et al.* applied similar concept for photo cropping that exploited sparse auto-encoder to discover the composition basis from a database containing well-composed images. Differ from [42], Wang's method organized the learning and inference in a cascade manner to achieve higher efficiency. By considering that perspective effect is one of the most commonly used techniques for photography, Zhou *et al.* developed a hierarchical segmentation method integrating photometric cues with perspective geometric cue to detect the dominant vanishing point in the image, which was employed for image re-composition or cropping [44].

Recently, thanks to the development of DNN, the research of image cropping tends to utilize deep learning approaches. By imitating the process of professional photographic, Chen *et al.* proposed a ranking CNN to harvest unambiguous pairwise aesthetic ranking examples on the web and applied this network to find the optimal cropping result from many candidate regions [45]. Instead of generating the attention map for cropping, Kao *et al.* proposed to use aesthetic map, which was extracted via a CNN, and gradient energy map to accomplish the image cropping task, by learning the composition rules through a SVM classifier [46]. In [47], Guo *et al.* designed a cascaded cropping regression (CCR) approach to crop the image, where a deep CNN was applied to extract features from images and the cropping areas were predicted by the proposed CCR algorithm. Inspired by human's decision making, Li *et al.*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.3029882, IEEE Transactions on Multimedia

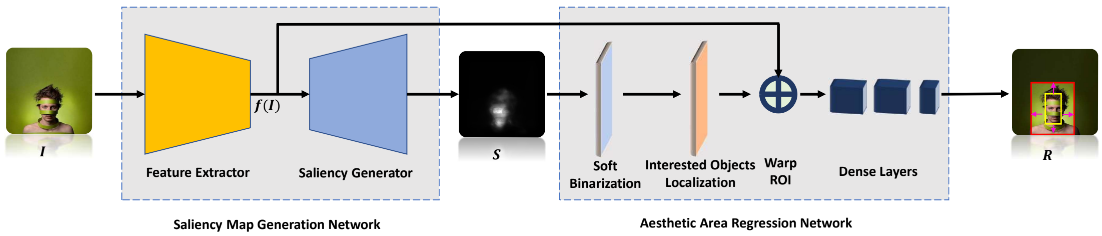


Fig. 1. Architecture of the proposed saliency map detection and aesthetic area regression network.

designed a weakly supervised aesthetic aware reinforcement learning framework to address the problem of image cropping, where the photo was initially cropped and repeatedly updated based on the current observation and the historical experience [2]. In [8], Lu *et al.* proposed a regression network based cropping method, which mapped initial detected saliency rectangle to a cropping area with high aesthetics quality. Unlike the conventional photo cropping method that only produced a single output, in [48], Wei *et al.* proposed a system that returned multiple cropping outputs based on a teacher-student framework. In this framework, the teacher was trained to evaluate candidate anchor boxes, and the scores from the teacher were used to supervise the training of student, a view proposal net. The combination of these two networks effectively improve the cropping performance. The interest readers can refer [49] for more comprehensive surveys.

III. Proposed Approach

A. Motivation & System Overview

By studying the procedure of professional photography, we can see that the theme is firstly determined by the photographer prior to other actions. To express this theme, the objects along with the compatible backgrounds are selected subsequently. Once the main objects contained in the photo are given, the other parameters for the photography, such as exposure time, composition, colors, etc., will be set for the final shooting.

Based on this observation, the process of image cropping to obtain the high aesthetic quality can be decomposed into two steps: detection of the interested objects $\mathcal{S}$ in the image $\mathcal{I}$ and prediction of the aesthetic areas $\mathcal{R}$ of the image based on the objects of interest $\mathcal{S}$. This process can be formally expressed as:

$$P(\mathcal{R},\mathcal{S}|\mathcal{I}) = P(\mathcal{S}|\mathcal{I}) \cdot P(\mathcal{R}|\mathcal{S},\mathcal{I}), \tag{1}$$

where $\mathcal{S} = \{S_{i,j} | i \times j \in |\mathcal{I}|\}$ denotes the interested objects of the image, $|\mathcal{I}|$ means the number of pixels in the image, and $\mathcal{S}_{i,j} \in \{0,1\}$ represents whether a given pixel belongs to the objects of interest. $P(\mathcal{S}|\mathcal{I})$ is the probability of interested objects $\mathcal{S}$ given an image $\mathcal{I}$, and $P(\mathcal{R}|\mathcal{S},\mathcal{I})$, which reveals the hidden relationship between the interested objects and the final cropping region, denotes the probability of $\mathcal{R}$ with respect to the image $\mathcal{I}$ and the detected objects of interest $\mathcal{S}$.

Thus, the aesthetic region of an image can be obtained if $P(\mathcal{R},\mathcal{S}|\mathcal{I})$ is calculated. Hence, a probabilistic model based cropping system can be designed whose parameters can be expressed as $\boldsymbol{\Theta}$, and the image cropping task can be considered as the maximum likelihood (ML) estimation problem for a given training image $\mathcal{I}_i$, along with its ground truths $\mathcal{S}_i$ and $\mathcal{R}_i$:

$$\begin{aligned}
\boldsymbol{\Theta} &= \arg\max_{\boldsymbol{\Theta}} \sum_{i=1}^{N} P(\mathcal{R}^{(i)},\mathcal{S}^{(i)}|\mathcal{I}^{(i)};\boldsymbol{\Theta}) \\
&= \arg\max_{\boldsymbol{\Theta}} \sum_{k=1}^{N} P(\mathcal{S}^{(k)}|\mathcal{I}^{(k)};\boldsymbol{\Theta}_s) \cdot P(\mathcal{R}^{(k)}|\mathcal{S}^{(k)},\mathcal{I}^{(k)};\boldsymbol{\Theta}_r) \\
&= \arg\max_{\boldsymbol{\Theta}} \sum_{k=1}^{N} \big(\log P(\mathcal{S}^{(k)}|\mathcal{I}^{(k)};\boldsymbol{\Theta}_s) \\
&\quad + \log P(\mathcal{R}^{(k)}|\mathcal{S}^{(k)},\mathcal{I}^{(k)};\boldsymbol{\Theta}_r)\big),
\end{aligned} \tag{2}$$

where superscript k indicates the index of training sample and ground truth, N is the total number of training samples, and $\boldsymbol{\Theta} = [\boldsymbol{\Theta}_s, \boldsymbol{\Theta}_r]^T$ are the parameters of the model.

Based on this analysis, we design an end-to-end DNN based image cropping system that follows the probability framework as described in Eq. 2. In the proposed cropping system, two main components are constructed, where the saliency map generation network $H(\mathcal{I};\boldsymbol{\Theta}_s)$ in the Fig. 1 is served to predict $\mathcal{S}$ given image $\mathcal{I}$. And aesthetic area regression network $G(\mathcal{I},\mathcal{S};\boldsymbol{\Theta}_r)$ containing the proposed IOL layer, ROI warping pooling layer and fully connected layers is used as a regressor to produce final cropping outputs $\mathcal{R}$ based on $\mathcal{I}$ and $\mathcal{S}$. These two components are corresponding to the photographer's actions of the objects decision and final cropping areas selection.

Thus, maximizing the Eq. 2 is equivalent to minimizing the loss L_{total} of the neural network:

$$\begin{aligned}
\boldsymbol{\Theta} &= \arg\min_{\boldsymbol{\Theta}} \mathcal{L}_{total} \\
&= \frac{1}{N}\sum_{k=1}^{N} \arg\min_{\boldsymbol{\Theta}} \Big(\mathcal{L}_s(\hat{\mathcal{S}}^{(k)},\mathcal{S}^{(k)}) + \lambda\mathcal{L}_r(\hat{\mathcal{R}}^{(k)},\mathcal{R}^{(k)})\Big),
\end{aligned} \tag{3}$$

where $\mathcal{L}_s(\cdot)$ represents the loss from the inconsistency between predicted $\hat{\mathcal{S}}^{(k)}$ by the saliency map detection network and ground truth $\mathcal{S}^{(k)}$ of the images $\mathcal{I}^{(k)}$, $\mathcal{L}_r(\cdot)$ is the loss caused by the difference between predicted aesthetic region $\hat{\mathcal{R}}^{(k)}$, and the ground truth region $\mathcal{R}^{(k)}$, and λ is the weight controlling the influence from these two networks and we use $\lambda = 1$ in this work.

As can be seen from the Fig. 1, unlike the conventional image cropping methods that explicitly or implicitly generate and evaluate multiple candidate cropping regions, the proposed system takes the input image to extract the interested object

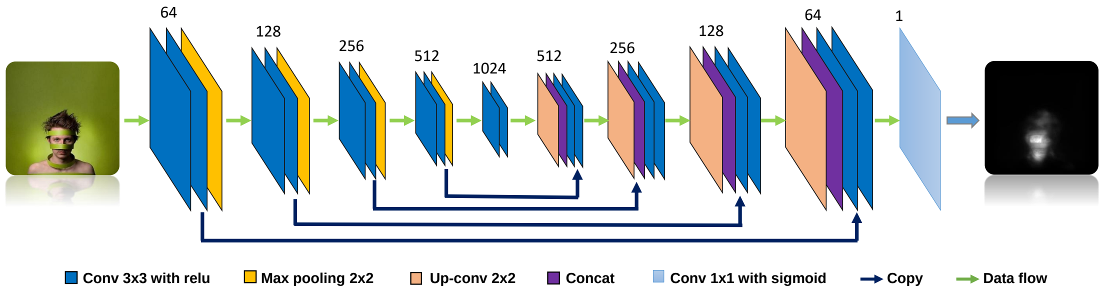


Fig. 2. The U-Shaped network implemented in this work for saliency map detection.

region and maps this single area to the final output region by regression network directly. Thus, in this framework the data flows through the network only once without extensively assessing multiple candidates, which highly improves the efficiency and maintains the accuracy in the meantime.

B. Saliency Map Generation Network

Saliency map detection aims at predicting visually interested objects in an image that attract human attention. In the proposed system, we adopt a modified U-shaped network to produce the salicency map. As a variant of widely used fully convolutonal encoder-decoder, U-shaped network is originally designed for semantic segmentation on biomedical images [50]. It merges feature maps from convolutional layers to deconvolutional layers gradually during the upsampling procedure. Thus, different types of features are preserved for the semantic labeling task.

Particularly, in our implementation, the encoder for the U-shaped network is composed by four fundamental blocks, where every two convolutional layers followed by a max pooling layer are stacked to form the basic block. Similarly, a decoder is constructed by four basic blocks where every two deconvolutional layers and a upsampling layer are used. For each fundamental block in the encoder, its feature maps are copied and concatenated directly to the corresponding block in the decoder with the same size of feature dimensions. Thus, the encoder is employed to extract features for the image and the saliency map is produced based on the decoder. The detailed structure of the U-shaped network we implemented is illustrated in Fig. 2.

C. Aesthetic Area Regression Network

Based on the image feature obtained through feature extractor and saliency map detected by the saliency generator, the relation between the interested objects and the area with high aesthetic quality can be learned through the proposed IOL layer and regression layers, which are described in the following subsections with details.

1) Soft Binarization Layer (SBL): To make our cropping system less sensitive to the presence of outliers in the saliency map, we introduce a function $\rho(x;\sigma)$ to enhance the quality of interested objects in saliency map, which is defined by:

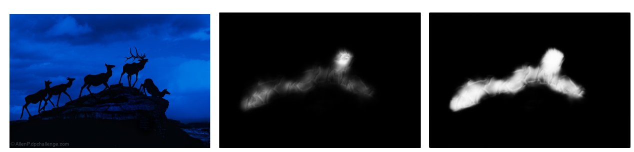
$$\rho(x;\sigma) = \frac{x^2}{x^2+\sigma^2}. \tag{4}$$

By selecting proper scale parameter σ, $\rho(x;\sigma)$ function maps small value of pixel in saliency map to 0, and saliency map will be saturated to 1 with larger pixel value. In Fig. 3, we demonstrate a sample saliency map and its corresponding enhanced version, from which we can observe that the difference between the interested objects in the image and the background is enlarged and they can be distinguished with minimum efforts.

As the derivative of $\rho(x;\sigma)$ function is calculated by:

$$\frac{\partial\rho(x;\sigma)}{\partial x} = \frac{2x\sigma^2}{(x^2+\sigma^2)^2}, \tag{5}$$

this operation can be easily integrated into the proposed neural network pipeline without blocking the backpropagation. And in our work, σ is empirically set as 0.01.



(a) (b) (c)

Fig. 3. Sample saliency map and enhanced interested object image for color image with high aesthetic scores. (a) The original high quality image. (b) The corresponding saliency map. (c) Enhanced saliency map by using function $\rho(x;\sigma)$, where the interested objects are easily distinguished from backgrounds.

2) Interested Object Localization (IOL) Layer: Based on the obtained soft binarization saliency map $\mathcal{S}$ that shows each pixel's probability to be the interested objects, it is necessary to model the $P(\mathcal{R}|\mathcal{S},\mathcal{I})$ to reveal the relation between the interested objects and the final cropping window. In order to represent this relation, it needs to extract the features of the interested objects first. To achieve this goal, the location of those interested objects needs to be determined. Generally, to locate the interested objects in the image, researchers extract the saliency map first and then search and locate the salient region based on it. However, most salient region localization methods use heuristic searching technique to scan all possible candidate regions, which are prohibitively slow. Even by using many speed up algorithms to reduce the searching space [32], [34], [51], [6], [52], those approaches are not differentiable which are infeasible in an end-to-end image cropping pipeline to allow the backpropagation. Thus, in this work, we propose an IOL layer that can effectively detect areas with interested objects in the image and is differentiable for end-to-end training.

Inspired by the mean shift algorithm that is used to locate and track the face regions in videos [53], a region generation algorithm is proposed in this work to perform the interested object region creation.

Given a saliency map $\mathcal{S}$ extracted by U-shaped network, the center of mass (c_x, c_y) for this map can be calculated by:

$$c_x = \frac{M_{10}}{M_{00}}, \quad c_y = \frac{M_{01}}{M_{00}},$$

and the standard deviation for the center of mass are obtained according to:

$$\sigma_x = \sqrt{\frac{M_{20}}{M_{00}} - c_x^2}, \quad \sigma_y = \sqrt{\frac{M_{02}}{M_{00}} - c_y^2},$$

where moments M_{00}, M_{01}, M_{10}, M_{20} and M_{02} are calculated based on:

$$M_{00} = \sum_{i,j} S_{i,j} \tag{6}$$

$$M_{10} = \sum_{i,j} i \cdot S_{i,j}, \quad M_{01} = \sum_{i,j} j \cdot S_{i,j} \tag{7}$$

$$M_{20} = \sum_{i,j} i^2 \cdot S_{i,j}, \quad M_{02} = \sum_{i,j} j^2 \cdot S_{i,j}. \tag{8}$$

Therefore, a region that includes the energy of the saliency map can be defined through its top-left corner (x_{tl}^s, y_{tl}^s) and bottom-right corner (x_{br}^s, y_{br}^s) by using a Gaussian-like window:

$$(x_{br}^s, y_{br}^s) = (c_x + \gamma\sigma_x, c_y + \gamma\sigma_y) \tag{9}$$

and

$$(x_{tl}^s, y_{tl}^s) = (c_x - \gamma\sigma_x, c_y - \gamma\sigma_y), \tag{10}$$

where γ is a hyper-parameter controlling the amount of energy contained in the window and maintaining the integrity of interested objects in the image. In this work, $\gamma = 3.0$ is applied to include over 99% energy from the interested objects in the image.

In Fig. 4, the examples for areas of interested object obtained by the IOL layer with different γ are illustrated. From these figures we can see that the IOL layers with $\gamma = 1.5$ can only cover partial of interested objects in the image. And when $\gamma = 3.0$, most areas of interested objects can be included whereas the background of the image is still excluded.

To allow backpropagation of the loss pass through this region generation layer, the gradient of the coordinates for interested object region w.r.t $\mathcal{S}$ can be defined. For coordinate x of bottom-right corner, the partial derivative is given by:

$$\frac{\partial x_{br}^s}{\partial S_{i,j}} = \frac{\partial c_x}{\partial S_{i,j}} + \gamma \frac{\partial \sigma_x}{\partial S_{i,j}}, \tag{11}$$

where $\frac{\partial c_x}{\partial S_{i,j}}$ and $\frac{\partial \sigma_x}{\partial S_{i,j}}$ are further calculated based on following equations:

$$\begin{aligned}\frac{\partial c_x}{\partial S_{i,j}} &= \frac{1}{M_{00}}\frac{\partial M_{10}}{\partial S_{i,j}} - \frac{M_{10}}{M_{00}^2}\frac{\partial M_{00}}{\partial S_{i,j}} \\ &= \frac{i}{M_{00}} - \frac{M_{10}}{M_{00}^2}\end{aligned} \tag{12}$$

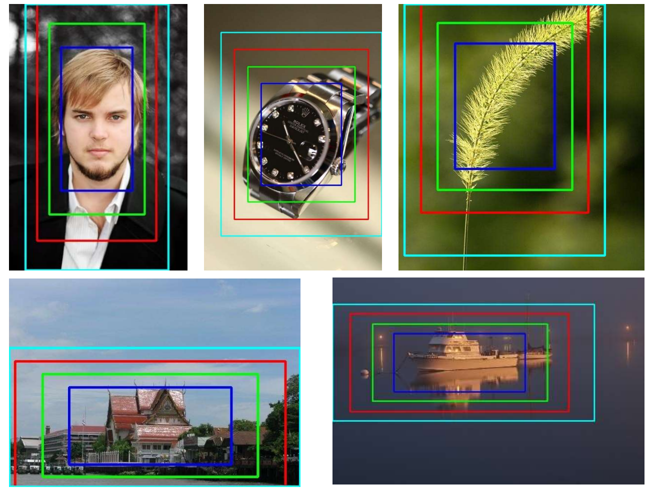


Fig. 4. The regions extracted by IOL layer with different γ. Blue box corresponds to $\gamma = 1.5$, green box is the area extracted by using $\gamma = 2.0$, red box is for $\gamma = 2.5$ and cyan box means the region obtained by $\gamma = 3.0$.

and

$$\begin{aligned}\frac{\partial \sigma_x}{\partial S_{i,j}} &= \frac{1}{2\sqrt{\frac{M_{20}}{M_{00}} - c_x^2}} \\ &\quad \times \left(\frac{1}{M_{00}}\frac{\partial M_{20}}{\partial S_{i,j}} - \frac{M_{20}}{M_{00}^2}\frac{\partial M_{00}}{\partial S_{i,j}} - 2c_x\frac{\partial c_x}{\partial S_{i,j}}\right) \\ &= \frac{1}{2\sqrt{\frac{M_{20}}{M_{00}} - c_x^2}} \\ &\quad \times \left\{\frac{i^2}{M_{00}} - \frac{M_{20}}{M_{00}^2} - 2c_x\left(\frac{i}{M_{00}} - \frac{M_{10}}{M_{00}^2}\right)\right\}\end{aligned} \tag{13}$$

and similar partial derivatives can be applied for $\frac{\partial y_{br}^s}{\partial S_{i,j}}$, $\frac{\partial x_{tl}^s}{\partial S_{i,j}}$ and $\frac{\partial y_{tl}^s}{\partial S_{i,j}}$.

This provides the proposed IOL layer a mechanism that allows loss gradients to flow back to the input of network to update the model's parameters $\boldsymbol{\Theta}$.

3) Dense Layers: Although the interested object region produced by the IOL layer contains most significant objects in the image, it is mostly far from having high aesthetic quality. So, based on the observation that the professional photographer tends to adjust the scene area for final shooting according to the interested objects, and the discovery that *"one may still roughly infer the extent of an object if only the middle of the object is visible"* [51], three fully connected layers are implemented to map the interested object region to the eventual cropping window with high visual quality based on its feature.

In our implementation, the region of interest (RoI) warping pooling layer followed by fully connected layers is used to estimate final cropping areas. The RoI warping pooling layer is proposed in [54], which takes two inputs: the coordinates of predicted interested object region and the feature maps generated from the bottle layer of U-shaped network. Prior to feeding into the RoI warping pooling layer, the coordinates of predicted interested object region are reduced 16 times to match the size of feature maps from bottle layer in U-shaped

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.3029882, IEEE Transactions on Multimedia

network. In the RoI warping pooling layer, only features from interested object region are extracted, which are consequently passed to two fully connected layers with ReLU activation, whose sizes are 2048 and 1024, respectively. The last layer of this regression network is a fully connected layer who has 4 units with linear activation function, which predicts the four coefficients defined by Eq. 16 and Eq. 17.

4) Aesthetic Area Representation: To represent the relation between the detected interested object region and areas with high aesthetic qualities, we use the approach described in [8]. Given a detected interested object region, whose size is $w^s \times h^s$, if its corresponding high aesthetic quality image's size is $w^a \times h^a$, and their top-left and bottom-right corners are (x_{tl}^s, y_{tl}^s), (x_{br}^s, y_{br}^s), (x_{tl}^a, y_{tl}^a) and (x_{br}^a, y_{br}^a), respectively, the offsets between the corners of these two rectangles $R\big((x_{tl}^s, y_{tl}^s), (x_{br}^s, y_{br}^s)\big)$ and $R\big((x_{tl}^a, y_{tl}^a), (x_{br}^a, y_{br}^a)\big)$ can be represented as:

$$(\Delta x_t, \Delta y_t) = (x_{tl}^s, y_{tl}^s) - (x_{tl}^a, y_{tl}^a) \tag{14}$$
$$(\Delta x_b, \Delta y_b) = (x_{br}^a, y_{br}^a) - (x_{br}^s, y_{br}^s). \tag{15}$$

Hence, the height and width of these two rectangles can be expressed as:

$$h^a = h^s + \Delta y_t + \Delta y_b = h^s + \alpha_t \cdot h^a + \alpha_b \cdot h^a \tag{16}$$

and

$$w^a = w^s + \Delta x_t + \Delta x_b = w^s + \beta_t \cdot w^a + \beta_b \cdot w^a, \tag{17}$$

where $\mathcal{O} = [\alpha_t, \alpha_b, \beta_t, \beta_b]$ are four coefficients.

In our implementation, the above coefficients $\mathcal{O} = [\alpha_t, \alpha_b, \beta_t, \beta_b]$ are used to represent the final aesthetic area and can be learned through a neural network.

During the testing, the corner coordinates (x_{tl}^s, y_{tl}^s), (x_{br}^s, y_{br}^s) of the interested object region of input image and four coefficients $\mathcal{O} = [\alpha_t, \alpha_b, \beta_t, \beta_b]$ are predicted by the proposed end-to-end network. Then, the width and height of final aesthetic region can be expressed as follows:

$$h^a = \frac{y_{br}^s - y_{tl}^s}{1 - \alpha_t - \alpha_b} \tag{18}$$
$$w^a = \frac{x_{br}^s - x_{tl}^s}{1 - \beta_t - \beta_b} \tag{19}$$

Thus, the coordinates of top-left and bottom-right corners of aesthetic area can be calculated by:

$$x_{tl}^a = x_{tl}^s - \beta_t \cdot w^a \qquad y_{tl}^a = y_{tl}^s - \alpha_t \cdot h^a$$
$$x_{br}^a = x_{br}^s + \beta_b \cdot w^a \qquad y_{br}^a = x_{br}^s + \alpha_b \cdot h^a$$

D. Loss Functions for the Cropping System

As introduced in subsection III-A, the total loss of the proposed neural network based cropping system is given by:

$$\mathcal{L}_{total} = \frac{1}{N} \sum_{k=1}^{N} \big(\mathcal{L}_s(\cdot) + \lambda \mathcal{L}_r(\cdot)\big) \tag{20}$$

where $\mathcal{L}_s(\cdot)$ is the loss from saliency map detection network and $\mathcal{L}_r(\cdot)$ is the loss from aesthetic regression network, N means the total training number, and λ is the weight controlling the influence from these two networks.

To train the U-shaped based network $H(\mathcal{I}, \mathbf{\Theta}_s)$, the binary cross-entropy of each pixel is calculated:

$$\begin{aligned} H\big(\mathcal{I}_{i,j}; \mathbf{\Theta}_s = (\mathbf{W}_s, \mathbf{b}_s)\big) \\ &= -S_{i,j} \log p(\mathcal{I}_{i,j}; (\mathbf{W}_s, \mathbf{b}_s)) \\ &- (1 - S_{i,j}) \log \big(1 - p(\mathcal{I}_{i,j}; (\mathbf{W}_s, \mathbf{b}_s))\big), \end{aligned} \tag{21}$$

where $[\mathbf{W}_s, \mathbf{b}_s]$ are weights of U-shaped saliency map detection network, $p(\mathcal{I}_{i,j}; (\mathbf{W}_s, \mathbf{b}_s))$ stands for the predicted confidence for the interested objects of each pixel, and $\hat{\mathcal{S}}_{i,j} = p\big(\mathcal{I}_{i,j}; (\mathbf{W}_s, \mathbf{b}_s)\big)$ holds for the detected saliency map $\hat{\mathcal{S}}$.

Thereafter, the loss for a given image $\mathcal{I}^{(k)}$ can be expressed as:

$$\begin{aligned} \mathcal{L}_s(\hat{\mathcal{S}}^{(k)}, \mathcal{S}^{(k)}) &= \mathcal{L}_s(\mathbf{W}_s, \mathbf{b}_s) \\ &= \sum_{\mathcal{I}_{i,j}^{(k)}} H\left(\mathcal{I}_{i,j}^{(k)}; (\mathbf{W}_s, \mathbf{b}_s)\right), \end{aligned} \tag{22}$$

where superscript k is the index of the training sample.

And as described in section III-C4, unlike other image cropping methods that train a ranker or classifier to evaluate the cropping areas' aesthetic quality by using training samples with high/low qualities, the proposed aesthetic region regression network uses a regressor to predict the cropping window, where only features from high aesthetic images are required and learned. Thus, in our training phase for the regression network, the interested object region $\mathcal{R}\big((x_{tr}^s, y_{tr}^s), (x_{bl}^s, y_{bl}^s)\big)$ for high aesthetic quality image is firstly detected by IOL layer. Then, the region of original high quality image $\mathcal{R}\big((x_{tr}^a, y_{tr}^a), (x_{bl}^a, y_{bl}^a)\big)$ is used to calculate the offsets coefficients $\mathcal{O} = [\alpha_t, \alpha_b, \beta_t, \beta_b]$, where $(x_{tr}^a, y_{tr}^a) = (0, 0)$, $(x_{bl}^a, y_{bl}^a) = (w^a, h^a)$ and $w^a \times h^a$ is high quality image's size. And these offsets coefficients are used to supervise the training of the proposed regression network, where $L2$ loss is applied according to:

$$\begin{aligned} \mathcal{L}_r(\hat{\mathcal{R}}^{(k)}, \mathcal{R}^{(k)}) &= \mathcal{L}_r(\mathbf{W}_r, \mathbf{b}_r) \\ &= \left\|\hat{\mathcal{O}}^{(k)} - \mathcal{O}^{(k)}\right\|^2 \end{aligned} \tag{23}$$

where $[\mathbf{W}_r, \mathbf{b}_r]$ are system weights for aesthetic regression network, $\mathcal{O}^{(k)}$ is the ground truth of offsets coefficients of an image $\mathcal{I}^{(k)}$ and $\hat{\mathcal{O}}^{(k)}$ is the corresponding predicted offsets coefficients.

IV. Experiments

A. Databases and Evaluation Protocol

We conducted our experiments on the following four databases.

1) Training database: In this experiment, the AVA database [55] was used for training. The AVA database, which was originally designed for aesthetic visual analysis, gathered more than 250,000 images from www.dpchallenge.com. Each image in AVA set contains plenty of meta-data, including multiple aesthetic scores from reviewers, semantic labels for over 60 categories, etc. In this work, we utilized AVA database to train the proposed end-to-end image cropping network, where only images whose average aesthetic scores were greater than or equal to 6 were selected for training, which resulted in a

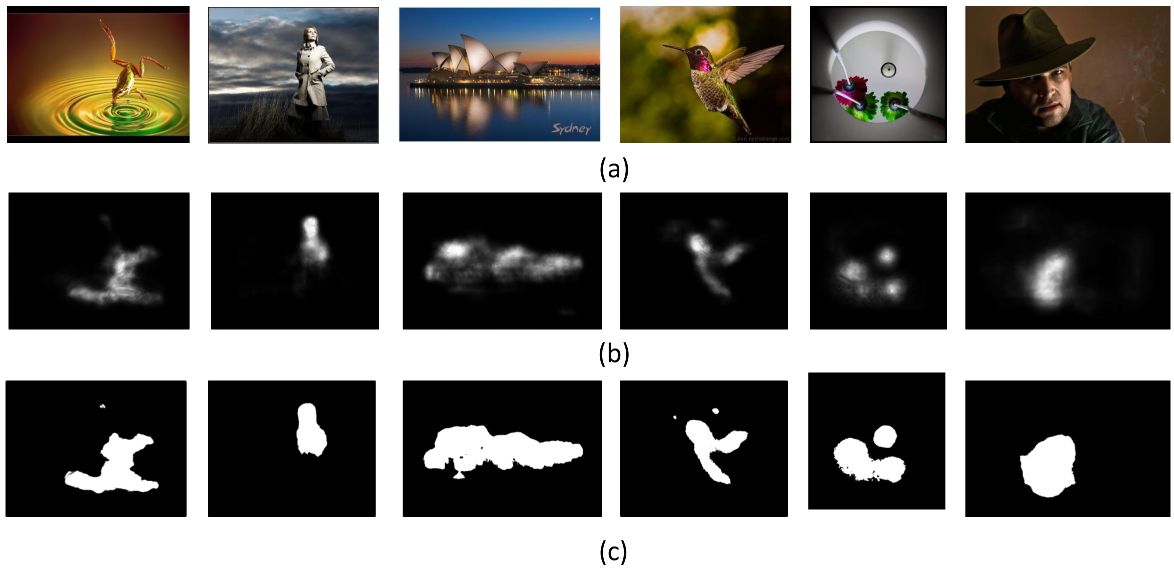


Fig. 5. Sample images along with their ground truths from AVA database. (a) Sample images with high aesthetic scores from AVA database. (b) Corresponding saliency maps for sample images. (c) Binarized interested object image for AVA samples where threshold is 0.12.

training set with $50,189$ high qualities images. Sample images from AVA database can be found in Figure 5 (a).

However, the AVA database was originally designed for aesthetic evaluation and only aesthetic scores were provided as the ground truths for each image. So in order to train the proposed neural networks with this database, the synthetic ground truths of interested object image and offsets of final crop window w.r.t. the interested object region for each image were produced initially. The preprocessing details for the training database is described in section IV-B accordingly.

2) Test databases: In our experiments, three public databases were applied for evaluation purpose.

The FCD database [56] was constructed to facilitate the aesthetic cropping task, where thousands images were collected from Flickr and cleaned by annotators. For each cleaned image, the cropping area was labeled by professional photographers and validated by multiple professional annotators who had passed Human Intelligence Tasks qualification test. Only those images that were ranked as preferable by at least 4 professional annotators were selected in the final cropping database. In our experiments, 334 samples were applied for evaluation purpose among this database.

FLMS database [3] collected 500 images from Flickr and the best cropping areas of each image were manually annotated by 10 experienced editors. In this work, we used FLMS database to evaluate the cropping performance.

Furthermore, to measure the proposed image cropping method, CUHK-ICD database [29] was employed. In this database, 950 images were captured by amateur photographers but cropped by 3 professional editors. All images in this database were used for evaluation in our work.

To quantitatively evaluate the cropping performance, the intersection over union (IoU) and boundary displacement error (BDE) were employed, where IoU is defined as:

$$IoU = \frac{A' \cap \hat{A}}{A' \cup \hat{A}} \tag{24}$$

and BDE is defined by:

$$BDE = \sum_{k=1}^{4} \left\| B'_k - \hat{B}_k \right\| / 4. \tag{25}$$

Here, A' means the ground truth of the cropping area, $\hat{A}$ represents the predicted cropping region, and B'_k and $\hat{B}_k$ are the normalized boundary coordinates for ground truth and predicted crop windows, respectively.

B. Neural Networks Training

Because the proposed image cropping system contained two main components conceptually, the parameters' search space is large with the joint training of entire network, which causes the low efficiency and unstable training. So, in this work, a corresponding three-stage training scheme was applied, where the U-shaped saliency map detection network $H(\mathcal{I}, \mathbf{\Theta}_s)$ and the regression network $G(\mathcal{I}, \mathbf{\Theta}_r)$ were trained sequentially and the entire network was fine-tuned afterwords.

To train the U-shaped network, the images from AVA database were employed, where $50,189$ images with their synthetic saliency maps were fed into the network for training. The synthetic saliency maps of AVA database were obtained by using method in [57], where an existing single branched DNN model was applied to detect the saliency maps of each image for AVA. Based on the obtained saliency map, the binarized image can be calculated by using a simple thresholding approach with empirical threshold, which was employed to guide the training of the proposed U-shaped network. In Figure 5 (b) and 5 (c) , the corresponding salieny maps and binarized interested objects for the sample images from AVA database are shown.

In this experiment, SGD optimization scheme was applied and the training rate was fixed to 1×10^{-4} for 4 epochs.

Once the U-shape network was learned, the obtained weights were locked for the second stage training of regression network. To train this regression network, the same training images with high qualities from AVA database were fed into U-shaped network to create saliency maps, from which an interested object region can be estimated based on the proposed IOL layer subsequently. Then, the coordinates of this interested object region were passed to RoI warping layer, where the corresponding features from U-shaped network were extracted and sent to the following fully connected layers, as illustrated in Figure 1.

The pre-calculated ground truths for offsets, which were obtained based on the method described in subsection III-D, were used to guide the training of regression network to predict the offset between interested object region and the final cropping rectangle. We used SGD optimizer with learning rate of 1×10^{-4} for 6 epochs in this training stage.

Finally, we used training images from AVA database along with the synthetic binarized saliency maps and pre-calculated offsets ground truths to fine-tune the entire network from end to end. The SGD optimizer was used in this stage with learning rate of 1×10^{-5} for 2 epochs and the U-shaped saliency map detection network and aesthetic area regression network have the same loss weight.

In this three-stage training phase, the input images were resized so that the shorter side of the image was 224 but the original aspect ratio was maintained.

C. Results Evaluation & Analysis

1) Comparison with the state-of-the-art approaches: To analyze the performance of the proposed end-to-end image cropping model, we compared the proposed method with other state-of-the-art cropping approaches, which were used as our baselines.

In table I, the IoUs and BDEs of the proposed cropping system and other state-of-the-arts cropping approaches on three public datasets are demonstrated, where * denotes weakly supervised cropping methods that do not use bounding boxes from annotated cropping datasets for training. As can be seen from this table, our cropping method obtained better cropping performances than any other approach on FLMS set. On FCD dataset, the proposed method achieved best performance among weakly supervised cropping approaches. And for CUHK-ICD database, the proposed method had competitive IoU and BDE performance on this evaluation set, which shows the effectiveness of the proposed cropping method.

In Fig. 6, multiple cropping results along with the corresponding detected saliency maps from the evaluation sets are demonstrated, where red boxes represent the optimized cropping window predicted by the proposed system and the green boxes show the detected IOLs based on Eq. 9 and Eq. 10 for saliency maps. From these images, we can see that the cropped images obtain better composition and aspect ratio than the original images, especially for those amateur captured low quality images.

2) Ablation test: To investigate the effectiveness of the proposed soft binarization layer (SBL), one cropping system was training where the SBL was removed. We illustrated and compared the cropping systems with/without SBL in the last two row of table I. From these numbers, we can observe that the IoU on CUHK-ICD dataset for the cropping system with SBL is higher than its counterpart without SBL for more than 5.0 on average. And the cropping results by the system with SBL on other test sets are also superior than the system without SBL. From these results, we can see that SBL can effectively help the cropping system to filter the noises and find the interested objects more accurately.

In the proposed image cropping framework, the U-shape based saliency map generation network can be easily replaced by other state-of-the-art saliency detection modules. Thus, in our experiments, we re-trained the SALICON saliency detection network, which was introduced in [57] and applied to generate the synthetic ground truth for AVA database in section IV-A, to detect the saliency maps for test images and consequently feed them into the IOL layer and aesthetic area regression network to produce the final cropping window. In table II, the overall cropping performances by combining the SALICON saliency map detection network and the proposed aesthetic area regression network are listed, where we can see it provides similar cropping results compared with the U-shape based saliency detection network, which shows the generalization capability of the proposed framework.

Need to note that in the ablation test, in order to avoid the size of feature maps extracted by SALICON saliency map detection network being too small, the input images of the neural networks were resized to ensure the shorter side of the image was 512 with the original aspect ratio.

By analyzing three tables and the structure of SALICON network and U-shaped network, it can be concluded that the cropping performance differences between these two saliency detection modules rely on the resolution of extracted features from these two networks. For the SALICON network, it applies the VGG-16 to extract down-sampled feature maps for images, which provides coarse details of the interested objects. But U-shaped saliency detection network extracts the feature map whose size is the same as the input image, that maintains more details of the interested objects. Therefore, for the cleaned high resolution images, such as photos from AVA database, U-shaped saliency detection network tends to extract more pleasant features of the interested objects in the image to help the cropping task. And for the noisy low quality images, SALICON network acts more like a noise suppressor to extract smoothed features to boost the cropping performance, as we observed from FCD database.

3) Investigation of image's size and aspect ratio: In many other research articles, it is claimed that the aesthetic quality of images is highly relied on the size or aspect ratio of the images [59], [60]. Thus, we carried out several experiments to investigate cropping performance with different image size and aspect ratio. In these experiments, we trained three models by keeping the original aspect ratio of the images but resizing the image till the shorter side of the image is 224, 384 or 512. Other three models were trained by resizing the image to square, whose size is 224×224, 384×384 or 512×512, respectively.

In table III, we list the IoUs and BDEs of different models

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.3029882, IEEE Transactions on Multimedia

TABLE I
THE CROPPING PERFORMANCE FOR DIFFERENT APPROACHES ON CUHK-ICD, FLMS AND FCD DATABASES.

Approach	CUHK-ICD						FLMS		FCD	
	Photographer1		Photographer2		Photographer3					
	IoU	BDE	IoU	BDE	IoU	BDE	IoU	BDE	IoU	BDE
*ATC [5]	0.605	0.108	0.628	0.100	0.641	0.095	0.720	0.063	0.58	0.10
*AIC [6]	0.469	0.142	0.494	0.131	0.512	0.123	0.640	0.075	0.47	0.13
*MPC [58]	0.603	0.106	0.582	0.112	0.608	0.110	0.410	N/A	N/A	N/A
*A2-RL [2]	0.802	0.052	0.796	0.054	0.790	0.054	0.820	N/A	N/A	N/A
*ABP-AA [1]	0.815	0.031	0.810	0.030	0.830	0.029	0.810	0.057	0.65	0.08
*VFN-SW [45] [2]	0.740	0.069	0.719	0.076	0.713	0.077	N/A	N/A	0.633	0.098
*Lu *et al.* [8]	0.827	0.032	0.816	0.035	0.805	0.036	0.843	0.029	0.659	0.062
VEN [48]	N/A	N/A	N/A	N/A	N/A	N/A	0.837	0.041	0.735	0.072
LCC [29]	0.748	0.066	0.728	0.072	0.732	0.071	0.630	N/A	N/A	N/A
*proposed w/o SBL	0.777	0.039	0.766	0.043	0.759	0.043	0.820	0.031	0.655	0.060
*Proposed w/ SBL	0.822	0.031	0.815	0.034	0.802	0.035	0.846	0.026	0.673	0.058

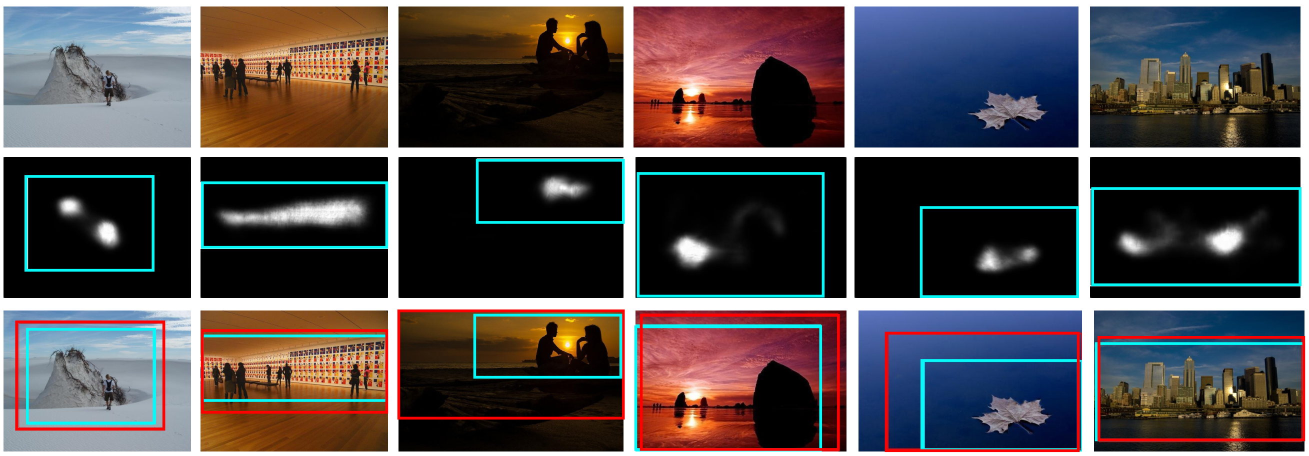


Fig. 6. Cropping rectangle produced by the proposed system.

TABLE II
THE CROPPING PERFORMANCE USING SALICON BASED SALIENCY DETECTION NETWORK ON CUHK-ICD, FLMS AND FCD DATABASES.

Method	CUHK-ICD						FLMS		FCD	
	Photographer1		Photographer2		Photographer3					
	IoU	BDE	IoU	BDE	IoU	BDE	IoU	BDE	IoU	BDE
Salicon + Regression	0.819	0.032	0.808	0.036	0.799	0.037	0.838	0.028	0.666	0.060
U-shaped + Regression	0.825	0.032	0.820	0.034	0.806	0.036	0.845	0.028	0.664	0.060

with various input image size on the three public test sets. From this table, it can be seen that both IoUs and BDEs for different input size of images have similar performances and no significant difference can be found between these models, which means the proposed image cropping model is insensitive to the size and aspect ratio of the input image.

By digging into table III, we observed that the overall IoU and BDE scores on CUHK-ICD database were getting better with larger input image size of neural networks, whilst the cropping performance was degraded on FCD database with larger input image size. The main reason of this phenomenon is that the images in the FCD database were collected from Flickr's website, containing more irrelevant background noises than the training database and other two evaluation databases. With a larger input size, more detailed features of images from FCD database, including non-interested background noises, can be discovered by the neural networks. But these features of noises cannot be effectively represented by the neural network which was trained based on the clean images from AVA database, and can be easily mis-represented as objects' features. This causes the proposed cropping method tending to generate larger crop windows to include more details when the input image size is big, which degrades the performance of the system on FCD database. But for the FLMS database, each test image had multiple annotated ground truths and the best cropping result was calculated using the ground truth which provided best performance. So, FLMS set is less sensitive to the neural network's input image size. With regard to CUHK-ICD database, it was constructed by high aesthetic quality images similar to AVA database, which in turn can be sufficiently embedded by the proposed cropping networks with larger input image size to attain better cropping performance.

Compromised by the cropping performance across three evaluation sets and the computation efficiency, it is preferable to resize the input image such that the shorter side is 224 for the proposed cropping approach.

4) Investigation of parameters σ and γ: From subsection III-C1, we can see that the quality of interested objects within the obtained saliency map can be enhanced by function $\rho(x;\sigma)$. To investigate the impact from scale parameter σ

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.3029882, IEEE Transactions on Multimedia

TABLE III
THE CROPPING PERFORMANCE FOR DIFFERENT ASPECT RATIO AND IMAGE SIZE ON CUHK-ICD, FLMS AND FCD DATABASES.

Input size	CUHK-ICD						FLMS		FCD	
	Photographer1		Photographer2		Photographer3					
	IoU	BDE	IoU	BDE	IoU	BDE	IoU	BDE	IoU	BDE
224×224	0.825	0.031	0.818	0.034	0.805	0.036	0.840	0.028	0.672	0.059
384×384	0.827	0.031	0.817	0.034	0.804	0.036	0.843	0.028	0.670	0.059
512×512	0.828	0.031	0.822	0.034	0.806	0.036	0.842	0.028	0.665	0.061
$\min(w, h) = 224$	0.822	0.031	0.815	0.034	0.802	0.035	0.846	0.026	0.673	0.058
$\min(w, h) = 384$	0.823	0.032	0.818	0.034	0.804	0.036	0.844	0.027	0.670	0.059
$\min(w, h) = 512$	0.825	0.032	0.820	0.034	0.806	0.036	0.845	0.028	0.664	0.060

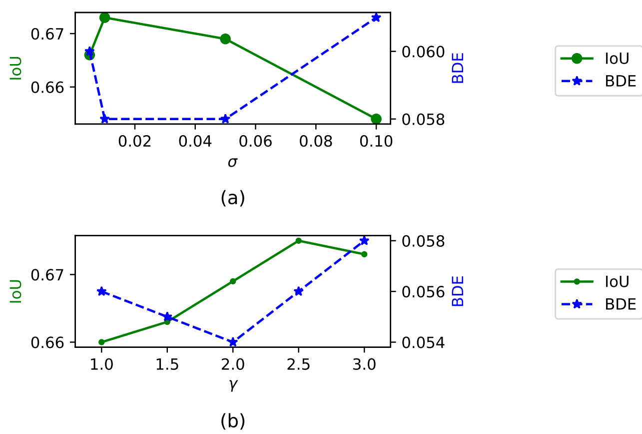


Fig. 7. The cropping performance for different σ and γ on FCD database.

of function $\rho(x;\sigma)$ for the cropping system, cropping performances (IoUs and BDEs) with different σs on test set FCD are shown in Fig. 7 (a), where function $\rho(x;\sigma)$ with $\sigma = \{0.005, 0.01, 0.05, 0.1\}$ are used. As can be seen from this figure, too small or large σ can cause lower cropping performance because more noises are introduced into IOLs or more interested objects are filtered out. Thus, a proper σ close to 0.01 provides us a better cropping result.

Similar to σ, the γ in Eq. 9 and Eq. 10 controls the amount of energy contained in the IOLs where larger γ causes more saliency areas but includes more background area also, and smaller γ results in the loss of integrity for salient object in IOLs. This effect can be seen in Fig. 7 (b), from which we can observe the cropping performances are getting degraded when γ is larger than 3.0 and smaller than 2.0 on the FCD test set.

5) Efficiency analysis: As one of the main contributions of this work is to use an end-to-end neural network to accomplish the image cropping task, without iteratively evaluating multiple candidates' aesthetic qualities, which has lower computational cost. So we measured time efficiency of the proposed system with different input size on the FLMS set, where the experiments were implemented with Keras on a server with Intel(R) Xeon(R) E5-2620 CPU @ 2.10GHz, 64Gb Memory and Nvidia 2080Ti GPU. We also compared our cropping method with other approaches w.r.t. speed, where the FPSs are shown in table IV. From this table, we notice that when the input image is resized to 224×224, the overall time for image cropping of our system is less than 20ms. Thus, the proposed system can reach over 50fps on average for real-time processing, which is much faster than other state-of-the-arts approaches and shows its high efficiency.

Furthermore, by comparing the time efficiency with the cropping method presented in [8] which is relied on a brute force search algorithm [6], the proposed cropping system is five times faster.

TABLE IV
THE TIME EFFICIENCY COMPARISON OF THE PROPOSED CROPPING SYSTEM WITH DIFFERENT SETTINGS AND OTHER METHODS.

Method	FPS
A2RL [2]	4
ABP-AA [1]	5
VFN [45]	0.5
Lu *et al.*[8]	10
proposed [224×224]	52
proposed [384×384]	29
proposed [512×512]	18
proposed [$\min(w, h) = 224$]	40
proposed [$\min(w, h) = 384$]	22
proposed [$\min(w, h) = 512$]	14

D. Subjective analysis

Because image's aesthetics is difficult to represent from the subjective perspective, such that different person might have different views for the same cropping results based on their tastes, education backgrounds, etc. So, in our work, a subjective comparative experiment was carried out.

In this experiment, 200 images were randomly collected from three (CUHK-ICD, FLMS and FCD) test sets. For each image, the proposed cropping method, along with the algorithms AIC [6], A2-RL [2] and VEN [48] was employed to obtain four cropping results. Then, 10 users were recruited, including 5 males and 5 females. All users had no prior knowledge of the experiment content and the databases. For each participant, the four cropping results of each test image were presented, where the order of the cropping images was randomized and the users were asked to vote the most pleasing one in terms of their aesthetics. Finally, 2000 votes from 10 participants were received and shown in figure 8.

As can be seen from this figure, the proposed method had gained most votes (792/2000) among four state-of-the-art cropping approaches, which shows the proposed system provided more pleasing cropping results than the other methods in respect to the aesthetics.

E. Case Study

To analyze the effects of different contents for the cropping performance, both success and failure cases in the evaluation are demonstrated.

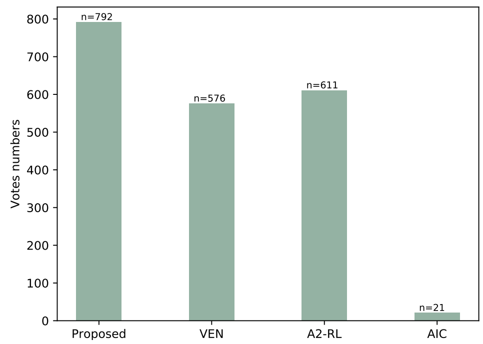
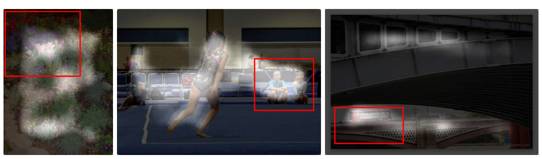


Fig. 8. Votes received from users for different state-of-the-art cropping methods.

(a) Failure examples from FLMS database.



(b) Failure cases from FCD database.

Fig. 9. Failure examples from the evaluation sets. (a) Failure samples from FLMS database, where red boxes are cropping windows by annotators. (b) Failure images from FCD database, where red boxes are ground truth and light areas are the detected saliency maps.

As shown in Fig. 6, when the interested object region are obtained from the saliency map successfully, the relations between the interested objects and the final cropping window can be learned by the proposed cropping system with sufficient training samples, where the area with the high aesthetic scores can also be inferred consequently.

Although the proposed image cropping approach works well on the majority of testing images, several failure cases can be found in the evaluation, which can be categorized into two types of errors broadly.

The first type of failures are mostly from the FLMS database, as shown in the Fig. 9(a), where only spurious texture regions exist in the image and it is hard to find enough salient pixels to determine the interested objects in the image. In our implementation, if no visual fixation is found, we use center areas that cover the 70% of entire image as our interested object region to feed into regression network to obtain the final cropping rectangle. The other type of failure cases can be seen in FCD database, where multiple interested objects are located by the saliency map detection network, as shown in Fig. 9(b), but only partial of these saliency area is included into the ground truth and most parts are missing, which causes the low IoUs and high BDEs.

V. Conclusion

In this paper, an end-to-end automatic image cropping system is proposed to learn the relationship between the interested objects and the areas with high aesthetic scores in an image through a DNN. Conceptually, the saliency map is initially detected by using a U-shaped neural network, which is then passed into a soft binarization layer to separate objects from the background. Based on this enhanced saliency map, an interested object region is determined by the proposed IOL layer, which is fed into a ROI warping pooling layer and following dense layers along with the features of the interested objects, to predict the optimal cropping region with high aesthetic scores.

As a weakly supervised cropping method, the proposed algorithm outperforms other weakly supervised state-of-the-art cropping methods w.r.t IoU and BDE metrics. Moreover, because the proposed approach finds the final cropping areas based on the hidden relationship between interested objects and areas with high aesthetics quality through neural networks, which avoids to iteratively evaluate multiple cropping candidates, high processing efficiency is achieved with 50 FPS.

Our future research will be exploring other cropping metric instead of IoU and BDE to measure the performance of aesthetics based cropping system.

References

[1] W. Wang, J. Shen, and H. Ling, "A deep network solution for attention and aesthetics aware photo cropping," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1531–1544, July 2019.

[2] D. Li, H. Wu, J. Zhang, and K. Huang, "A2-RL: Aesthetics aware reinforcement learning for image cropping," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8193–8201.

[3] 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 22Nd ACM International Conference on Multimedia*, 2014, pp. 1105–1108.

[4] M. Nishiyama, T. Okabe, Y. Sato, and I. Sato, "Sensation-based photo cropping," in *Proceedings of the 17th ACM International Conference on Multimedia*, ser. MM '09, 2009, pp. 669–672.

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

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

[7] W. Wang and J. Shen, "Deep cropping via attention box prediction and aesthetics assessment," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2205–2213.

[8] P. Lu, H. Zhang, X. Peng, and X. Peng, "Aesthetic guided deep regression network for image cropping," *Signal Processing: Image Communication*, vol. 77, pp. 1 – 10, 2019.

[9] X. Wang, H. Ma, and X. Chen, "Salient object detection via fast r-cnn and low-level cues," in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 1042–1046.

[10] N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 678–686.

[11] Y. Ji, H. Zhang, and Q. J. Wu], "Salient object detection via multi-scale attention cnn," *Neurocomputing*, vol. 322, pp. 130 – 140, 2018.

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

[13] L. Itti and C. Koch, "Comparison of feature combination strategies for saliency-based visual attention systems," in *Human Vision and Electronic Imaging IV*, vol. 3644, 1999, pp. 473 – 482.

[14] ——, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research*, vol. 40, no. 10, pp. 1489 – 1506, 2000.

[15] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proceedings of the 19th International Conference on Neural Information Processing Systems*, ser. NIPS'06, 2006, pp. 545–552.

[16] 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.

[17] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *2009 IEEE 12th International Conference on Computer Vision*, 2009, pp. 2106–2113.

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

[19] P. Moon and D. E. Spencer, "Geometric formulation of classical color harmony," *J. Opt. Soc. Am.*, vol. 34, no. 1, pp. 46–59, Jan 1944.

[20] T. Grill and M. Scanlon, *Photographic composition*. New York, N.Y.: Amphot, 1983.

[21] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2006, pp. 419–426.

[22] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *Computer Vision – ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds., 2006, pp. 288–301.

[23] L. Zhang, Y. Gao, R. Zimmermann, Q. Tian, and X. Li, "Fusion of multichannel local and global structural cues for photo aesthetics evaluation," *IEEE Transactions on Image Processing*, vol. 23, no. 3, pp. 1419–1429, March 2014.

[24] X. Tang, W. Luo, and X. Wang, "Content-based photo quality assessment," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1930–1943, 2013.

[25] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *2011 International Conference on Computer Vision*, 2011, pp. 1784–1791.

[26] Y. Kao, R. He, and K. Huang, "Deep aesthetic quality assessment with semantic information," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1482–1495, 2017.

[27] P. Lu, X. Peng, X. Zhu, and R. Li, "An EL-LDA based general color harmony model for photo aesthetics assessment," *Signal Processing*, vol. 120, pp. 731 – 745, 2016.

[28] Z. Liu, Z. Wang, Y. Yao, L. Zhang, and L. Shao, "Deep active learning with contaminated tags for image aesthetics assessment," *IEEE Transactions on Image Processing*, in press.

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

[30] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Computational intelligence and neuroscience*, vol. 2018, p. 7068349, 2018.

[31] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27 – 48, 2016.

[32] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[33] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[34] R. Girshick, "Fast R-CNN," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.

[35] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[36] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6517–6525.

[37] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., 2016, pp. 21–37.

[38] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollr, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2999–3007.

[39] 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.

[40] L. Zhang, M. Song, Y. Yang, Q. Zhao, C. Zhao, and N. Sebe, "Weakly supervised photo cropping," *IEEE Transactions on Multimedia*, vol. 16, no. 1, pp. 94–107, 2014.

[41] L. Zhang, Y. Gao, R. Ji, Y. Xia, Q. Dai, and X. Li, "Actively learning human gaze shifting paths for semantics-aware photo cropping," *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 2235–2245, 2014.

[42] A. Samii, R. Měch, and Z. Lin, "Data-driven automatic cropping using semantic composition search," *Comput. Graph. Forum*, vol. 34, no. 1, pp. 141–151, 2015.

[43] P. Wang, Z. Lin, and R. Mech, "Learning an aesthetic photo cropping cascade," in *2015 IEEE Winter Conference on Applications of Computer Vision*, 2015, pp. 448–455.

[44] Z. Zhou, S. He, J. Li, and J. Z. Wang, "Modeling perspective effects in photographic composition," in *Proceedings of the 23rd ACM International Conference on Multimedia*, 2015, pp. 301–310.

[45] Y.-L. Chen, J. Klopp, M. Sun, S.-Y. Chien, and K.-L. Ma, "Learning to compose with professional photographs on the web," in *Proceedings of the 2017 ACM on Multimedia Conference*, 2017, pp. 37–45.

[46] Y. Kao, R. He, and K. Huang, "Automatic image cropping with aesthetic map and gradient energy map," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 1982–1986.

[47] G. Guo, H. Wang, C. Shen, Y. Yan, and H. M. Liao, "Automatic image cropping for visual aesthetic enhancement using deep neural networks and cascaded regression," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2073–2085, 2018.

[48] Z. Wei, J. Zhang, X. Shen, Z. Lin, R. Mech, M. Hoai, and D. Samaras, "Good view hunting: Learning photo composition from dense view pairs," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[49] M. B. Islam, W. Lai-Kuan, and W. Chee-Onn, "A survey of aesthetics-driven image recomposition," *Multimedia Tools Appl.*, vol. 76, no. 7, pp. 9517–9542, 2017.

[50] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., 2015, pp. 234–241.

[51] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.

[52] N. Doulamis, A. Doulamis, D. Kalogeras, and S. Kollias, "Low bit-rate coding of image sequences using adaptive regions of interest," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 8, pp. 928–934, 1998.

[53] G. R. Bradski, "Computer vision face tracking for use in a perceptual user interface," 1998.

[54] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3150–3158.

[55] N. Murray, L. Marchesotti, and F. Perronnin, "Ava: A large-scale database for aesthetic visual analysis," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2408–2415.

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

[57] X. Huang, C. Shen, X. Boix, and Q. Zhao, "Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 262–270.

[58] J. Park, J. Y. Lee, Y. W. Tai, and I. S. Kweon, "Modeling photo composition and its application to photo re-arrangement," in *2012 19th IEEE International Conference on Image Processing*, 2012, pp. 2741–2744.

[59] S. Ma, J. Liu, and C. W. Chen, "A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 722–731.

[60] L. Mai, H. Jin, and F. Liu, "Composition-preserving deep photo aesthetics assessment," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 497–506.