Photo Cropping via Deep Reinforcement Learning

Yaqing Zhang
School of Digital Media and Design Arts
Beijing University of Posts and Telecommunications
Beijing, China
zhangyaq@bupt.edu.cn

Xueming Li
Beijing Key Laboratory of Network System and Network Culture
Beijing University of Posts and Telecommunications
Beijing, China
lixm@bupt.edu.cn

Xuewei Li
Department of Information and Communication Engineering
Beijing University of Posts and Telecommunications
Beijing, China
lixuewei@bupt.edu.cn

***Abstract*—Automatic image cropping aims at changing the composition of images to improve the aesthetic quality of images. It can provide professional advice for image editors and save time. Most of the existing automatic image cropping methods are based on specific features such as aesthetic features or salient features. These methods adopt sliding window mechanism to generate numerous cropping candidates, and then select the final results based on these specific features. It is very time-consuming and can only produce cropping results of a limited aspect ratio. In the face of these situations, a DLRL (deep learning framework combined with reinforcement learning) framework is proposed for image cropping, which only uses the basic features of the image for cropping without producing numerous candidate windows. Moreover, cropping step by step is more in line with the process of image cropping by people using Photoshop or other software. Experiments show that the proposed method can save a lot of time and improve cropping efficiency. The method proposed achieves the state-of-art performance in the open Flickr Cropping Dataset and CUHK Image Cropping Dataset.**

Keywords—image cropping, reinforcement learning, deep learning

I. INTRODUCTION

Image cropping is the most common and important task in image editing, which aims to extract well-composed regions from ill-composed images to improve the aesthetic quality of images. An excellent automatic image cropping algorithm can provide professional suggestions for image editors and save a lot of time.

In the field of automatic image cropping, many novel methods have been proposed [1,2,3,5,7,8]. Most cropping method based on specific features [6,1,7] adopt sliding window mechanism to generate numerous cropping windows. As shown in Fig. 1, these methods generally can be divided into the following three steps: 1) Use sliding window method to extract numerous cropping candidates on the input image, 2) Extract specific features of each cropping candidate, such as aesthetic features or salient features, 3) According to the evaluation of extracted features, select the best results from these cropping candidates. All these methods have achieved good performance, but it is impossible to produce cropping results with arbitrary aspect ratios due to the limitation of sliding window mechanism. Additionally, thousands of cropping windows are generated each time, which not only increases the time complexity but also increases the space complexity. It has very high requirements on hardware. Meanwhile, these cropping methods do not conform to the process of human cropping by using Photoshop and other software. When people use software to crop images, they will browse the whole image at first and adjust the cropping region gradually on the basis of the whole image until they find the most satisfactory area.

Based on the above investigation and inspired by the process of human cropping, we propose a DLRL (deep learning framework combined with reinforcement learning) framework for automatic image cropping. As shown in Fig. 2, the proposed method extracts the overall basic features of the image at first corresponding to the process of human browsing image. Then cropping regions are adjusted step by step on the basis of the whole image until the results can not be improved. Through the above process, the proposed method will not produce numerous cropping candidates and can achieve satisfactory performance by using less training data. Especially, the method proposed in this paper can finish cropping within 5 steps on average, which is very efficient.

Fig. 1. The process of most image cropping methods.

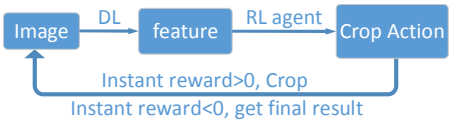


Fig. 2. The process of the method proposed in this paper. Image features are extracted through DL (deep learning part) and sent into RL (reinforcement learning part) to obtain a cropping action. If the action can improve the result, then cropping is performed. Otherwise, the current image will be output as the final result.

II. RELATED WORK

Automatic image cropping is to change the composition to improve the aesthetic quality of images. There are two main types of automatic image cropping methods: attention-based methods and aesthetic-based methods. The goal of attention-based methods [9,15,16,17] is to find the most visually prominent area in the original image. These methods usually select cropping windows according to the attention score or the salience of the object. These methods are usually suitable for removing the unimportant content of the image and sometimes fail to produce visually pleasing results due to the lack of aesthetic consideration of the image. Aesthetic image cropping method [18,6] aims to find the most satisfactory cropping window from the original image. Because these methods use aesthetic quality as a criterion, they use almost the same characteristics as aesthetic quality assessment. C. Fang et al. [6] and M. Nishiyama et al. [18] used aesthetic quality classifier to distinguish the quality of candidate windows. Y.L. Chen et al. [12] used RankSVM, and J. Klopp et al. [7] used RankNet to rate each candidate window. J. Yan et al. proposed a change-based method[4], which compared the original images with cropped images in order to discard distracting areas and obtain high-quality images.

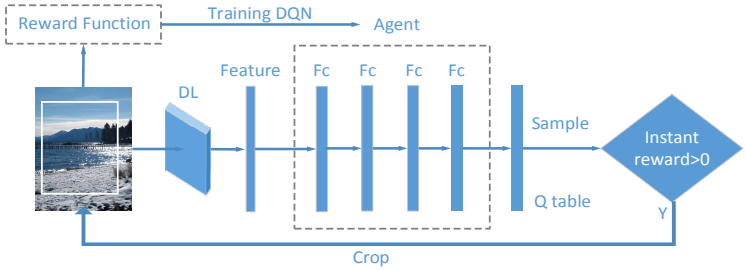


Fig. 3. Overall structure of the proposed method.

At present, most feature-based automatic image cropping methods [6,7,1] still rely on sliding window method to obtain cropping candidates. They seek the region with the highest attention score or aesthetic score from numerous cropping candidates. W. Wang et al. [2] made an optimization in this aspect. They firstly selected an initial cropping window and randomly generated cropping candidates around the initial window, which narrows the range of sliding window. But they did not break the limitation of the above sliding window method. The use of sliding window method limits the cropping results with arbitrary size. More importantly, these methods need to generate numerous cropping candidates to complete the cropping process. People will not randomly crop numerous candidate results on the image and then choose a satisfactory result from all these candidates. It is worth mentioning that the idea of gradual cropping is proposed in [5,13].

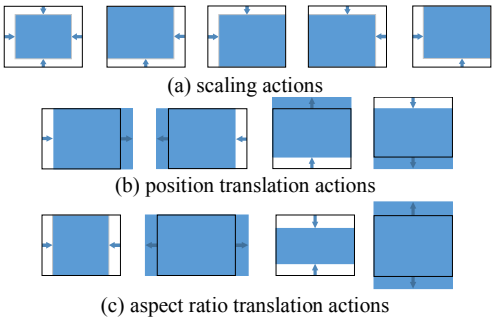
In this paper, a DLRL model is proposed for automatic image cropping. Due to the characteristics of reinforcement learning, the proposed method executes cropping step by step, which conforms to the process of human cropping. And DLRL does not need to produce numerous cropping candidates. In theory, cropping results with arbitrary aspect ratio can be generated. This method only uses the basic features of the image and achieves satisfactory cropping performance.

III. Our Approach

The DLRL (deep learning framework combined with reinforcement learning) method proposed in this paper improves the cropping results gradually, which is in line with the decision-making process of people cropping images. The overall network structure is shown in Fig. 3. As shown in Fig. 3, input image $I(t)$ is the image result after t steps of cropping, where t refers to the number of cropping steps. DL(deep learning part) is used to extract the feature of $I(t)$. Then the feature is sent to the agent as $s(t)$. According to the selection policy and Q table, agent selects an action A. If the reward of action A is greater than zero, the cropping action A will be executed, and $I(t+1)$ is obtained. Repeat the above process. Otherwise, the cropping ends. In the process of gradual cropping, the agent interacts with the environment (image). According to the state (feature) of the current environment, the agent selects the appropriate action from the action space. During the training, the reward for executing the action will be recorded so that the next time the agent can select a better action. Repeat this process and the optimal results will be obtained. In this part, we firstly introduce the action space and deep learning part of DLRL. Then the reward and agent in reinforcement training are introduced. Finally, we introduce the overall structure and the details of the network.

A. Action Space

There are 13 actions in the action space, which can be divided into three categories: scaling actions, position translation actions, and aspect ratio translation actions. The size, position and shape of the cropping window can be adjusted in these three categories respectively, and the corresponding number of actions is 5, 4, and 4 respectively. See Fig. 4 for specific information. The frame line is the boundary of the current image, and the dark part is the result after adjustment. The arrow indicates the direction of scaling or displacement. Take the first image in Fig. 4(a) as an example. It means that based on the current image, the aspect ratio remains unchanged and the cropping region is reduced inwards. The definition of these actions is similar to [19], except that we select a smaller adjustment ratio -- 0.025, so that the adjustment accuracy is higher and the optimal results are more likely to be obtained.



(a) scaling actions

(b) position translation actions

(c) aspect ratio translation actions

Fig. 4. Details of the action space.

B. Deep Learning Part

The method proposed in this paper is based on the basic features of the image for cropping, so the extraction of features is very important for cropping results. The prototype of DL part is VGGNet[14]. As we all know, VGGNet itself is used for classification and not suitable for cropping tasks. And the training data for VGGNet are extremely complex. So we fine tune the fc6 layer of VGGNet on Flickr Cropping Dataset (FCD)[12] and retain the original settings in conv1-conv5 layer. The network structure is shown in Table 1.

TABLE I. Last Three Layers of Deep Learning Part

Layer		Setting
conv5	conv5_1	filter_size=[3,3], filter_num=512, stride=[1,1,1,1]
	conv5_2	filter_size=[3,3], filter_num=512, stride=[1,1,1,1]
	conv5_3	filter_size=[3,3], filter_num=512, stride=[1,1,1,1]
max pool		kernel=[1,2,2,1], stride=[1,2,2,1]
fc6		num_in=49*512, num_out=4096

TABLE II. TEST RESULTS ON CUHK-ICD[4]

Method	Annotation 1			Annotation 2			Annotation 3		
	Avg Iou	*Avg Disp*	*α-recall*	*Avg Iou*	*Avg Disp*	*α-recall*	*Avg Iou*	*Avg Disp*	*α-recall*
VGG	0.3168	0.2268	1.16%	0.3196	0.2245	1.16%	0.3163	0.2274	0.63%
VGG + RL	0.6848	0.0776	48.42%	0.6809	0.0784	46.94%	0.6817	0.0777	48.63%
VGG_finetune+RL	**0.8206**	**0.0474**	**75.84%**	**0.8267**	**0.0466**	**70.46%**	**0.8047**	**0.0510**	**65.82%**

C. Reward and Agent

The agent used in DLRL includes four fully-connected layers, and the output size of the layers are 4096, 4096, 512, 13. After receiving DL extracted features, the agent estimates each action and then choose a cropping action according to the policy Ω. The policy Ω is determined with an ϵ-greedy algorithm. The ϵ-greedy algorithm randomly samples actions with a probability of ϵ and selects actions with the highest action score with a probability of $1-\epsilon$. Action score $Q(S(t),A)$ is the expected cumulative reward obtained by selecting action A. It use γ-discount cumulative reward to compute, and the formula is:

$$Q(S(t),A) = R(S(t),A) + \gamma \max_a Q(S',A) \quad (1)$$

$$R(S(t),A) = r(t) - r(t-1) \quad (2)$$

where $r(t)$ is the overlap degree between the present cropping window and the ground truth window after t steps of cropping, which is represented by the commonly used cropping evaluation metric IoU. The specific calculation formula is shown in Equation (3). $R(S(t),A)$ is the instant reward obtained by selecting action A, which measures the improvement of step t. During training, the agent always chooses the action with the highest action score, which means $\epsilon = 0$ in selection policy. The cropping process is repeated until $R(S(t),A) \leq 0$. In other words, the cropping ends when the result can not be improved.

D. Reinforement Learning based Cropping

DL network extracts the basic features of the input image *I(t)* as the current state $s(t)$ and send it to the agent. The agent selects a cropping action A from the action space, according to Q table. If the instant reward $R(S(t),A) > 0$, the action will be performed. Then the image after cropping will be input to DL network as $I(t+1)$. Repeat this process. If $R(S(t),A) \leq 0$, no cropping action is performed, and the cropping ends. The final cropping result is the input of DL network at that time, which means $I(t-1)$.

We use Double q-learning method [21] to train DLRL on Flickr Cropping Dataset[12]. The total number of training images is 1299, and batch size is 4. The initial value of learning rate is set to 10^{-5}, and the minimum value is set to 10^{-8}. The learning rate decreases every 5000 iterations, decreasing by 0.96 times. The maximum value of instant reward is set to 0.5, and the minimum value is set to -0.5. The cumulative reward is calculated with γ =0.95.

IV. EXPERIMENTS

We tested DLRL on Flickr Cropping Dataset (FCD)[12] and CUHK Image Cropping Dataset (CUHK-ICD)[4]. FCD contains 332 test images, each with one cropping annotation. CUHK-ICD contains 950 test images, each with three cropping annotations. Note that none of the test data is seen in the training.

A. Evaluation Metrics

Three common evaluation metrics of automatic image cropping are adopted, including average intersection-over-union (IoU), α-recall[7] and average boundary displacement (Disp).

The IoU measures the overlap degree between the cropping results and ground truth cropping window. The average IoU is calculated as follows:

$$IoU_i = (S_i^c \cap S_i^g)/(S_i^c \cup S_i^g) \quad (3)$$

$$Avg\ IoU = 1/N \sum_{i=1}^{N} IoU_i \quad (4)$$

where S_i^c is the region after cropping of the image *i*, and S_i^g is the region of the ground truth cropping window of the image *i*. *N* is the total number of images.

α-recall measures the proportion of good cropping results in all cropping results, which can reflect the general cropping level. The calculation used in this paper is as follows:

$$1/N \sum_{i=1}^{N} coumt(IoU_i > 0.75) \quad (5)$$

where IoU_i is the same as defined in Equation (3).

Disp measures the average distance between each side of the cropping window and that of the ground truth cropping window. The average Disp is calculated as follows:

$$Disp_i = (\left|B_i^{l,c} - B_i^{l,g}\right|/width + \left|B_i^{r,c} - B_i^{r,g}\right|/width + \left|B_i^{u,c} - B_i^{u,g}\right|/height + \left|B_i^{b,c} - B_i^{b,g}\right|/height)/4 \quad (6)$$

$$Avg\ Disp = 1/N \sum_{i=1}^{N} Disp_i \quad (7)$$

where {l, r, u, b} corresponds to the left, right, top and bottom respectively. $B_i^{l,c}$ denotes the left boundary value of the cropping window of the image *i*, and $B_i^{l,g}$ denotes the left boundary value of the ground truth cropping window and so on. Note that the boundary distance in x direction and y direction is normalized by the width and height of the image respectively.

B. Experimental results and analysis

1) Experiments for vgg fine tune

VGGNet and VGG_finetune are compared on FCD and CUHK-ICD as well as before and after adding reinforcement learning. We train VGGNet with 50 epochs on the FCD with the output of 4 corresponding to 4 coordinates of the cropping window. And select the optimal model from 50 epochs as baseline VGG. VGG_finetune was the optimal model after

TABLE III. TEST RESULTS ON FCD[12]

Method	Avg Iou	Avg Disp	α-recall
VGG	0.2092	0.2542	0.0%
VGG + RL	0.5904	0.0973	24.40%
VGG_finetune+RL	**0.6738**	**0.0849**	**36.75%**

TABLE IV. TEST RESULTS OF LOSS ON CUHK-ICD[4]

Method	Annotation 1			Annotation 2			Annotation 3		
	Avg Iou	*Avg Disp*	*α-recall*	*Avg Iou*	*Avg Disp*	*α-recall*	*Avg Iou*	*Avg Disp*	*α-recall*
mse	0.7527	0.0647	67.93%	0.7296	0.0712	65.61%	0.7395	0.0679	65.93%
iou	**0.8206**	**0.0474**	75.84%	**0.8267**	**0.0466**	70.46%	**0.8047**	**0.0510**	65.82%
disp	0.7992	0.0526	**76.82%**	0.7813	0.0575	**71.35%**	0.7829	0.0559	**70.57%**

fine tuning the fc6 layer with 10 epochs on the FCD training set, as described above. The results are as follows: Table 2 shows the test results on CUHK-ICD and Table 3 shows the results on FCD. It can be seen from the experimental results that VGGNet itself is not suitable for the cropping task, and the result increases significantly after fine tuning. The addition of RL also improves the cropping quality impressively.

2) Experiments for loss

We conducted a comparative experiment of loss on CUHK-ICD and FCD, and the results are as shown in Table 4 and Table 5. Loss refers to the composition of $r(t)$ in Equation (2). Iou and disp correspond to Equation (4) and Equation (7) respectively. Mse is calculated as follows:

$$Mse = [(x_0^c - x_0^g)^2 + (x_1^c - x_1^g)^2 + (y_0^c - y_0^g)^2 + (y_1^c - y_1^g)^2]/4 \quad (8)$$

where $[x_0^c,\ x_1^c, y_0^c, y_1^c]$ denotes the coordinates of the cropping window , and $[x_0^g,\ x_1^g, y_0^g, y_1^g]$ denotes that of the ground truth cropping window.

TABLE V. TEST RESULTS OF LOSS ON FCD[12]

Method	Avg Iou	Avg Disp	α-recall
mse	0.6696	0.0856	34.94%
iou	**0.6738**	**0.0849**	**36.75%**
disp	0.6676	0.0871	36.36%

According to the experimental results, selecting iou to calculate $r(t)$ has the best cropping effect. So iou is used to calculate $r(t)$ in the final model.

C. Compare with other methods

1) Evaluation for Cropping Accuracy

To evaluate the cropping accuracy of DLRL, experiments are conducted on FCD and CUHK-ICD. The cropping methods chosen for comparison cover all the common kinds of automatic image cropping methods. eDN[9] is an attention-based cropping method. It selects the final cropping window by maximizing the difference of saliency score between the retained part and the discarded part. The optimal results given in [12] are selected for the comparison results. MNA-CNN[11] and RankSVM+FCD[12] are both aesthetics-based cropping methods. The results given in [7] are selected for comparison. VFN+SW[7] is an aesthetics-based cropping method. It generates numerous cropping candidates by sliding window method, and the optimal results are selected through aesthetic evaluation. The comparison results were selected from the original paper. ABP+AA[2] is a comprehensive evaluation method based on both aesthetics and saliency, and the optimal results in the original paper are selected for the comparison results. A2RL[13] is an aesthetics-based cropping method. It uses view finding network(VFN)[7] for aesthetic evaluation, and the cropping is carried out gradually via reinforcement learning. The results for comparison are the best ones given in the original paper. The specific experimental results are as follows: Table 6 shows the results on FCD and Table 7 shows the results on CUHK-ICD. It can be seen from the test results that the cropping quality of this method is satisfactory. Not only are the cropping results closer to the ground truth, but good cropping results also account for a higher proportion in the overall cropping results. Note that the Avg IoU and Avg Disp of VFN+SW[7] on FCD are both better than ours, and we will discuss this further in Section (2).

TABLE VI. CROPPING ACCURACY ON FCD[12]

Method	Avg Iou	Avg Disp	α-recall
eDN[9]	0.4857	0.1372	12.68%
MNA-CNN[11]	0.5042	0.1361	0.07%
RankSVM+FCD[12]	0.602	0.1057	18.10%
VFN+SW[7]	**0.6842**	0.0843	35.06%
ABP+AA[2]	0.65	**0.08**	--
A2RL[13]	0.6633	0.0892	--
DLRL	**0.6738**	**0.0849**	**36.75%**

2) Evaluation for Time Efficiency

We test the cropping efficiency on FCD with two metrics: the average cropping time (Avg Time) and the average cropping steps (Avg Steps), which are used to measure the time and steps taken to cut a single image on average. Meanwhile, the results of IoU and Disp are retained to illustrate the cropping quality. The test experiment is conducted on a single NVIDIA GeForce GTX 1080 Ti GPU with Intel(R) Core(TM) i7-6800k CPU. Since A2RL[13] is also based on reinforcement learning, we select A2RL for comparison. In the experiment, the upper limit of cropping steps in A2RL[13] is set to 20, which is the same as the original setting in [13]. VFN+SW[7] uses all the original settings in [7], and we regard the number of cropping candidates generated as the cropping steps. The experimental results are shown in Table 8.

TABLE VIII. TIME EFFICIENCY ON FCD[12]

Method	Avg Iou	Avg Disp	Avg Time/s	Avg Step
VFN+SW[7]	0.6441	0.0943	33.9350	1125
A2RL[13]	0.6635	0.0887	0.8536	17.1566
DLRL	**0.6738**	**0.0849**	**0.2964**	**2.997**

TABLE VII. CROPPING ACCURACY ON CUHK-ICD[4]

Method	Annotation 1			Annotation 2			Annotation 3		
	Avg Iou	*Avg Disp*	*α-recall*	*Avg Iou*	*Avg Disp*	*α-recall*	*Avg Iou*	*Avg Disp*	*α-recall*
eDN[9]	0.5535	0.1273	27.37%	0.5128	0.1419	20.11%	0.5257	0.1358	22.42%
MNA-CNN[11]	0.4693	0.1555	0.07%	0.4553	0.1615	0.06%	0.4610	0.1590	0.07%
RankSVM+FCD[12]	0.6683	0.0907	33.47%	0.6618	0.0932	32.11%	0.6483	0.0973	31.26%
VFN+SW[7]	0.7847	0.0581	59.79%	0.7763	0.0614	58.11%	0.7602	0.0653	54.84%
ABP+AA[2]	0.81	**0.031**	--	0.810	**0.030**	--	**0.830**	**0.029**	--
A2RL[13]	0.809	0.0524	--	0.7961	0.0535	--	0.7902	0.0535	--
DLRL	**0.8206**	0.0474	**75.84%**	**0.8267**	0.0466	70.46%	0.8047	0.0510	**65.82%**

From Table 8, we can find out that VFN+SW[7] with 1125 cropping candidates generated is still inferior to our method, which means that the results of VFN+SW[7] given in Table 6 need more than 1125 candidates and 33 seconds to outperform DLRL by less than 1%. According to all the experimental results and analysis, the method proposed has the best performance on automatic image cropping. It can complete cropping within 5 steps, with both efficiency and quality.

V. CONCLUSION

In this paper, a DLRL framework (deep learning framework combined with reinforcement learning) is proposed for automatic image cropping. The method crops step by step according to the basic features of the image, which conforms to the human cropping mode. And it does not need to generate numerous cropping candidates. Experimental results show that DLRL not only improves the efficiency of image cropping, but also achieves excellent cropping effect.

REFERENCES

[1] Y. Kao, R. He, and K. Huang. Automatic image cropping with aesthetic map and gradient energy map. In ICASSP,2017.

[2] 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, 2018.

[3] S. A. Esmaeili, B. Singh, and L. S. Davis. Fast-at: Fast automatic thumbnail generation using deep neural networks. In CVPR, 2017.

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

[5] E. Hong, J. Jeon, and S. Lee. Cnn based repeated cropping for photo composition enhancement. In CVPR workshop, 2017.

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

[7] Y.-L. Chen, J. Klopp, M. Sun, S.-Y. Chien, and K.-L. Ma. Learning to compose with professional photographs on the web. In ACM Multimedia, 2017

[8] L. Zhang, M. Song, Y. Yang, Q. Zhao, C. Zhao, and N. Sebe.Weakly supervised photo cropping. IEEE Transactions on Multimedia, 2014.

[9] E. Vig, M. Dorr, and D. Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2798–2805, 2014. 5, 6

[10] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In ECCV, 2016. 1, 2, 3, 5, 6

[11] L. Mai, H. Jin, and F. Liu. Composition-preserving deep photo aesthetics assessment. In CVPR, 2016. 1, 2, 3, 4, 5, 6,7

[12] Y.-L. Chen, T.-W. Huang, K.-H. Chang, Y.-C. Tsai, H.-T. Chen, and B.-Y. Chen. Quantitative analysis of automatic image cropping algorithms: A dataset and comparative study. In WACV, 2017. 2, 3, 4, 5, 6, 9

[13] D. Li, H. Wu, J. Zhang, and K. Huang, "A2-rl: Aesthetics aware reinforcement learning for image cropping," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp.8193–8201.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," International Conference on Learning Representations (ICLR), 2015.

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

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

[17] F. Stentiford. Attention based auto image cropping. In Workshop on Computational Attention and Applications, ICVS, 2007.

[18] M. Nishiyama, T. Okabe, Y. Sato, and I. Sato. Sensationbased photo cropping. In ACM Multimedia, 2009.

[19] Z. Jie, X. Liang, J. Feng, X. Jin, W. Lu, and S. Yan. Treestructured reinforcement learning for sequential object localization. In NIPS, 2016.

[20] K. He, X. Zhang, S. Ren and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904-1916, 1 Sept. 2015.

[21] H. Van Hasselt, A. Guez, and D. Silver. Deep reinforcement learning with double q-learning. In AAAI, pages 2094–2100, 2016. 5

[22] Henderson, Peter & Islam, Riashat & Bachman, Philip & Pineau, Joelle & Precup, Doina & Meger, David. (2017). Deep Reinforcement Learning that Matters.

[23] P. Wang, Z. Lin and R. Mech, "Learning an Aesthetic Photo Cropping Cascade," 2015 IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, 2015, pp. 448-455.

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

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

[26] Y. Deng, C. C. Loy, and X. Tang, "Image aesthetic assessment: An experimental survey," IEEE Signal Processing Magazine, vol. 34, no. 4, pp. 80–106, 2017.

[27] L. Mai, H. Jin, and F. Liu, "Composition-preserving deep photo aesthetics assessment," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 497–506.

[28] B. Cheng, B. Ni, S. Yan, and Q. Tian, "Learning to photograph," in Proceedings of the ACM International Conference on Multimedia, 2010, pp. 291–300.

[29] W. Luo, X. Wang, and X. Tang, "Content-based photo quality assessment," in Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 2206–2213.