Grid Anchor based Image Cropping: A New Benchmark and An Efficient Model

Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang, *Fellow, IEEE*

Abstract—Image cropping aims to improve the composition as well as aesthetic quality of an image by removing extraneous content from it. Most of the existing image cropping databases provide only one or several human-annotated bounding boxes as the groundtruths, which can hardly reflect the non-uniqueness and flexibility of image cropping in practice. The employed evaluation metrics such as intersection-over-union cannot reliably reflect the real performance of a cropping model, either. This work revisits the problem of image cropping, and presents a grid anchor based formulation by considering the special properties and requirements (e.g., local redundancy, content preservation, aspect ratio) of image cropping. Our formulation reduces the searching space of candidate crops from millions to no more than ninety. Consequently, a grid anchor based cropping benchmark is constructed, where all crops of each image are annotated and more reliable evaluation metrics are defined. To meet the practical demands of robust performance and high efficiency, we also design an effective and lightweight cropping model. By simultaneously considering the region of interest and region of discard, and leveraging multi-scale information, our model can robustly output visually pleasing crops for images of different scenes. With less than 2.5M parameters, our model runs at a speed of 200 FPS on one single GTX 1080Ti GPU and 12 FPS on one i7-6800K CPU. The code is available at: https://github.com/HuiZeng/Grid-Anchor-based-Image-Cropping-Pytorch.

Index Terms—Image cropping, photo cropping, image aesthetics, deep learning.

1 Introduction

Cropping is an important and widely used operation to improve the aesthetic quality of captured images. It aims to remove the extraneous contents of an image, change its aspect ratio and consequently improve its composition [1]. Cropping is a high-frequency need in photography but it is tedious when a large number of images are to be cropped. Therefore, automatic image cropping has been attracting much interest in both academia and industry in past decades [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13].

Early researches on image cropping mostly focused on cropping the major subject or important region of an image for small displays [2], [14] or generating image thumbnails [15], [16]. Attention scores or saliency values were the major considerations of these methods [17], [18]. With little consideration on the overall image composition, the attention-based methods may lead to visually unpleasing outputs [5]. Moreover, user study was employed as the major criteria to subjectively evaluate cropping performance, making it very difficult to objectively compare different methods.

Recently, several benchmark databases have been released for image cropping research [5], [6], [10]. On these databases, one or several bounding boxes were annotated by experienced human subjects as "groundtruth" crops for each image. Two objective metrics, namely intersection-over-union (IoU) and boundary displacement error (BDE) [19], were defined to evaluate the performance of image cropping models on these databases. These public benchmarks enable many researchers to develop and test their cropping models, significantly facilitating the research on automatic image cropping [5], [10], [11], [13], [20], [21], [22], [23], [24].

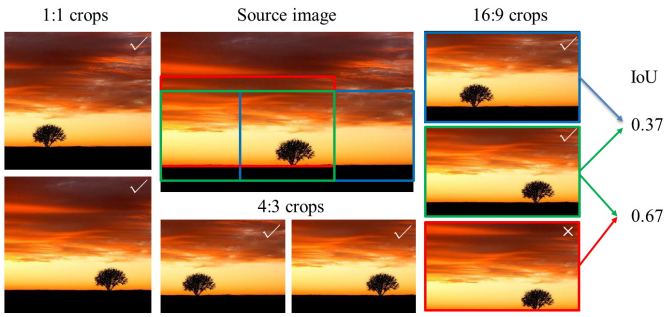


Fig. 1. The property of non-uniqueness of image cropping. Given a source image, many good crops (labeled with "√") can be obtained under different aspect ratios (e.g., 1:1, 4:3, 16:9). Even under the same aspect ratio, there are still multiple acceptable crops. Regarding the three crops with 16:9 aspect ratio, by taking the middle one as the groundtruth, the bottom one (a bad crop, labeled with "×") will have obviously larger IoU (intersection-over-union) than the top one but with worse aesthetic quality. This shows that IoU is not a reliable metric to evaluate cropping quality.

Though many efforts have been made, there are several intractable challenges caused by the special properties of image cropping. As illustrated in Fig. 1, image cropping is naturally a subjective and flexible task without unique solution. Good crops can vary significantly under different requirements of aspect ratio and/or resolution. Even under certain aspect ratio or resolution constraint, acceptable crops can also vary. Such a high degree of freedom makes the existing cropping databases, which have only one or several annotations, difficult to learn reliable and robust cropping models.

The commonly employed IoU and BDE metrics are unreliable to evaluate the performance of image cropping models either. Re-

- *H. Zeng, L. Li and L. Zhang are with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong. (email: {cshzeng, cslli, cslzhang}@comp.polyu.edu.hk).*
- *Z. Cao is with Camera Group of DJI Innovations Co., Ltd, Shenzhen, China (e-mail: zisheng.cao@dji.com).*

TABLE 1
IoU scores of recent representative works and the developed models in this work on two existing cropping benchmarks in comparison with two simplest baselines. Baseline_N simply calculates the IoU between the groundtruth and source image without cropping. Baseline_C crops the central part whose width and height are 0.9 time of the source image.

Method	ICDB [5]			FCDB [10]
	Set 1	Set 2	Set 3	
Yan *et al.* [5]	0.7487	0.7288	0.7322	–
Chen *et al.* [10]	0.6683	0.6618	0.6483	0.6020
Chen *et al.* [21]	0.7640	0.7529	0.7333	**0.6802**
Wang *et al.* [11]	0.8130	0.8060	0.8160	0.6500
Wang *et al.* [25]	0.8150	0.8100	**0.8300**	–
Li *et al.* [13]	0.8019	0.7961	0.7902	0.6633
Baseline_N	**0.8237**	**0.8299**	0.8079	0.6379
Baseline_C	0.7843	0.7599	0.7636	0.6647
GAIC (Mobile-V2)	0.8179	0.8150	0.8070	0.6735
GAIC (Shuffle-V2)	0.8199	0.8170	0.8050	0.6751

ferring to the three crops with 16:9 aspect ratio in Fig. 1, by taking the middle one as the groundtruth, the bottom one, which is a bad crop, has obviously larger IoU than the top one, which is a good crop. Such a problem can be more clearly observed from Table 1. By using IoU to evaluate the performance of recent works [5], [10], [11], [13], [21] on the existing cropping benchmarks ICDB [5] and FCDB [10], most of them have even worse performance than the two simplest baselines: no cropping (i.e., take the source image as cropping output, denoted by Baseline_N) and central crop (i.e., crop the central part whose width and height are 0.9 time of the source image, denoted by Baseline_C).

The special properties of image cropping make it a challenging task to train an effective and efficient cropping model. On one hand, since the annotation of image cropping (which requires good knowledge and experience in photography) is very expensive [10], existing cropping databases [5], [6], [10] provide only one or several annotated crops for about 1,000 source images. On the other hand, the searching space of image cropping is very huge, with millions of candidate crops for each image. Clearly, the amount of annotated data in current databases is insufficient to train a robust cropping model. The unreliable evaluation metrics further constrain the research progress on this topic.

In order to address the above issues, we reconsider the problem of image cropping and propose a new approach, namely grid anchor based image cropping, to accomplish this challenging task in a reliable and efficient manner. Our contributions are threefold.

- We propose a grid anchor based formulation for image cropping by considering the special properties and requirements (e.g., local redundancy, content preservation, aspect ratio) of this problem. Our formulation reduces the number of candidate crops from millions to no more than ninety, providing an effective solution to satisfy the practical requirements of image cropping.
- Based on our formulation, we construct a new image cropping database with exhaustive annotations for each source image. With a total of 106,860 annotated candidate crops and each crop annotated by 7 experienced human subjects, our database provides a good platform to learn robust image cropping models. We also define three new types of metrics which can more reliably evaluate the performance of learned cropping models than the IoU and BDE used in previous datasets.
- We design an effective and efficient image cropping model under the convolutional neural network (CNN) architecture. Specifically, our model first extracts multi-scale features from the input image and then models both the region of interest and region of discard to stably output a visually pleasing crop. Leveraging the recent advances in designing efficient CNN models [26], [27], our cropping model contains less than 2.5M parameters, and runs at a speed of up to 200 FPS on one single GTX 1080Ti and 12 FPS on CPU.

This paper extends our conference version [28] in four aspects. (1) More evaluation metrics are defined to evaluate more comprehensively the cropping models. (2) The feature extraction modules (VGG16 [29] and ResNet50 [30]) in the conference version are replaced by more efficient architectures (MobileNetV2 [26] and ShuffleNetV2 [27]), which significantly improve the efficiency of our cropping model without sacrificing the performance. (3) A multi-scale feature extraction architecture is designed to more effectively handle the images with varying scales of objects. (4) A set of effective data augmentation strategies are employed for learning photo composition, which further improve the performance of trained model. With all these improvements, our new model has much smaller size, much higher efficiency and much better cropping results.

2 Related Work

In this section, we summarize the existing image cropping datasets and evaluation metrics, representative image cropping methods and efforts made on improving cropping efficiency.

2.1 Image cropping datasets and evaluation metrics

Although the research of image cropping has been lasting for more than one decade, subjective assessment was employed as the major evaluation criteria for a long time because of the highly subjective nature and the expensive annotation cost of image cropping. Yan *et al.* [5] constructed the first cropping dataset, which consists of 950 images. Each image was manually cropped by three photographers. Contemporarily, Feng *et al.* [6] constructed a similar cropping dataset which contains 500 images with each image cropped by 10 expert users. Both datasets employed the IoU and BDE to evaluate the cropping performance. Unfortunately, the limited number of annotated crops is insufficient to learn a robust cropping model and the evaluation metrics are unreliable for performance evaluation. To obtain more annotations, Chen *et al.* [10] proposed a pairwise annotation strategy. They built a cropping dataset consisting of 1,743 images and 31,430 annotated pairs of subviews. Using a two-stage annotation protocol, Wei *et al.* [24] constructed a large scale comparative photo composition (CPC) database which can generate more than 1 million view pairs. Although the pairwise annotation strategy provides an efficient way to collect more training samples, the candidate crops in both datasets are either randomly generated or randomly selected from cropping results generated by previous cropping methods. These candidate crops are unable to provide more reliable and effective evaluation metrics for image cropping.

Different from the previous ones, our dataset is constructed under a new formulation of image cropping. Our dense annotations not only provide extensive information for training cropping model but also enable us to define new evaluation metrics to more reliably evaluate the cropping performance.

2.2 Image cropping methods

The existing image cropping methods can be divided into three categories according to their major drives.

Attention-driven methods. Earlier methods are mostly attention-driven, aiming to identify the major subject or the most informative region of an image. Most of them [2], [15], [16], [18] resort to a saliency detection algorithm (e.g. [31]) to get an attention map of an image, and search for a cropping window with the highest attention value. Some methods also employ face detection [32] or gaze interaction [17] to find the important region of an image. However, a crop with high attention value may not necessarily be aesthetically pleasing.

Aesthetic-driven methods. The aesthetic-driven methods improve the attention-based methods by emphasizing the overall aesthetic quality of images. These methods [5], [6], [32], [33], [34], [35], [36], [37] usually design a set of hand-crafted features to characterize the image aesthetic properties or composition rules. Some methods further deign quality measures [32], [35] to evaluate the quality of candidate crops, while some resort to training an aesthetic discriminator such as SVM [33], [34]. The release of two cropping databases [5], [6] further facilitates the training of discriminative cropping models. However, the handcrafted features are not strong enough to accurately predict the complicated image aesthetics [20].

Data-driven methods. Most recent methods are data-driven, which train an end-to-end CNN model for image cropping. Limited by the insufficient number of annotated training samples, many methods in this category [10], [11], [13], [20], [22], [23], [25] adopt a general aesthetic classifier trained from image aesthetic databases such as AVA [38] and CUHKPQ [39] to help cropping. However, a general aesthetic classifier trained on full images may not be able to reliably evaluate the crops within one image [21], [24]. An alternative strategy is to use pairwise learning to construct more training data [10], [21], [24].

Our method lies in the data-driven category with several advantages over the existing methods. First, we propose a new formulation for image cropping learning. Second, we constructed a much larger scale dataset with reliable annotations, which enables us to train more robust and accurate cropping models.

2.3 Image cropping efficiency

Efficiency is important for a practical image cropping system. Two types of efforts can be made to improve the efficiency: reducing the number of candidate crops and decreasing the computational complexity of cropping models. A brutal force sliding window search can easily result in million of candidates for each image. To reduce the number of candidate crops, Wang *et al.* [25] first detected the salient region of an image and then generated about 1,000 crops around the salient region. This strategy inevitably suffers from the same problem faced by attention-based methods: many useful background pixels are unnecessarily lost and the cropping results may not have the best composition. Wei *et al.* [24] employed the pre-defined 895 anchor boxes in the single shot detector (SSD) [40]. Again, the anchor boxes designed for object detection may not be the optimal choice for image cropping. Our new formulation carefully considers the special properties of image cropping and reduces the number of candidate crops to be no more than 90.

Regarding the model complexity, most recent cropping models are based on the AlexNet [10], [21] or VGG16 architecture [11], [20], [24], which are too heavy to be deployed on computational resource limited devices such as mobile phones and drones. Our cropping model embraces the latest advances in efficient CNN architecture design and it is much more lightweight and efficient than the previous models.

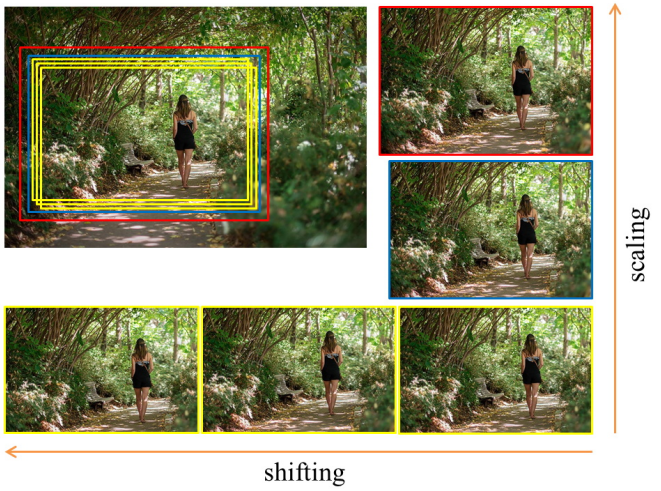


Fig. 2. The local redundancy of image cropping. Small local changes (e.g., shifting and/or scaling) on the cropping window of an acceptable crop (the bottom-right one) are very likely to output acceptable crops too.

3 Grid Anchor Based Image Cropping

As illustrated in Fig. 1, image cropping has a high degree of freedom. There is not a unique optimal crop for a given image. We consider two practical requirements of a good image cropping system. Firstly, a reliable cropping system should be able to return acceptable results for different settings (e.g., aspect ratio and resolution) rather than one single output. Secondly, the cropping system should be lightweight and efficient to run on resource limited devices. With these considerations, we propose a grid anchor based formulation for practical image cropping, and construct a new benchmark under this formulation.

3.1 Formulation

Given an image with spatial resolution $H \times W$, a candidate crop can be defined using its top-left corner (x_1, y_1) and bottom-right corner (x_2, y_2), where $1 \leq x_1 < x_2 \leq H$ and $1 \leq y_1 < y_2 \leq W$. It is easy to calculate that the number of candidate crops is $\frac{H(H-1)W(W-1)}{4}$, which is a huge number even for an image of size 100×100. Fortunately, by exploiting the following properties and requirements of image cropping, the searching space can be significantly reduced, making automatic image cropping a tractable problem.

Local redundancy: Image cropping is naturally a problem with local redundancy. As illustrated in Fig. 2, a set of similar and acceptable crops can be obtained in the neighborhood of a good crop by shifting and/or scaling the cropping widow. Intuitively, we can remove the redundant candidate crops by defining crops on image grid anchors rather than dense pixels. The proposed grid anchor based formulation is illustrated in Fig. 3. We construct an image grid with $M \times N$ bins on the original image, and define the corners (x_1, y_1) and (x_2, y_2) of one crop on the grid centers,

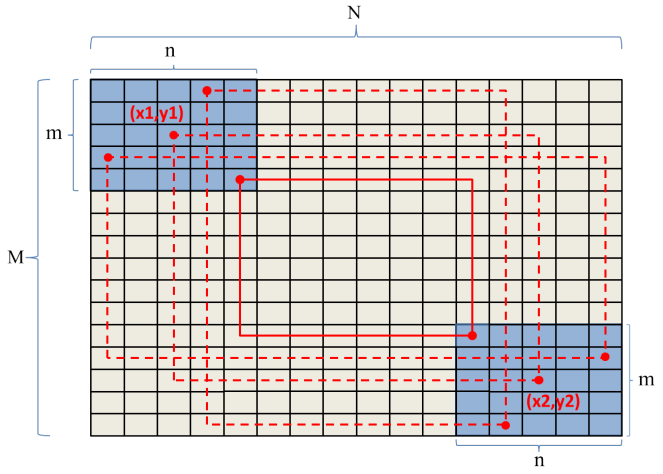


Fig. 3. Illustration of the grid anchor based formulation of image cropping. M and N are the numbers of bins for grid partition, while m and n define the adopted range of anchors for content preservation.

Fig. 4. The content preservation of image cropping. The small crop (b) misses the two persons, which are the key objects in the original image although itself has a good composition. With content preservation constraint, crop (a) will be generated to preserve as much useful information as possible.

which serve as the anchors to generate a representative crop in the neighborhood. Such a formulation largely reduces the number of candidate crops from $\frac{H(H-1)W(W-1)}{4}$ to $\frac{M(M-1)N(N-1)}{4}$, which can be several orders smaller.

Content preservation: Generally, a good crop should preserve the major content of the source image [6]. Otherwise, the cropped image may miss important information in the source image and misinterpret the photographer's purpose, resulting in unsatisfied outputs. An example is shown in Fig. 4. As can be seen, without the content preservation constraint, the output crop with good composition may miss the two persons in the scene, which are the key objects in the original image. Therefore, the cropping window should not be too small in order to avoid discarding too much the image content. To this end, we constrain the anchor points (x_1, y_1) and (x_2, y_2) of a crop into two regions with $m \times n$ bins on the top-left and bottom-right corners of the source image, respectively, as illustrated in Fig. 3. This further reduces the number of crops from $\frac{M(M-1)N(N-1)}{4}$ to m^2n^2.

The smallest possible crop (highlighted in red solid lines in Fig. 3) generated by the proposed scheme covers about $\frac{(M-2m+1)(N-2n+1)}{MN}$ grids of the source image, which may still be too small to preserve enough image content. We thus further constrain the area of potential crops to be no smaller than a certain proportion of the whole area of source image:

$$S_{crop} \geq \lambda S_{Image}, \tag{1}$$

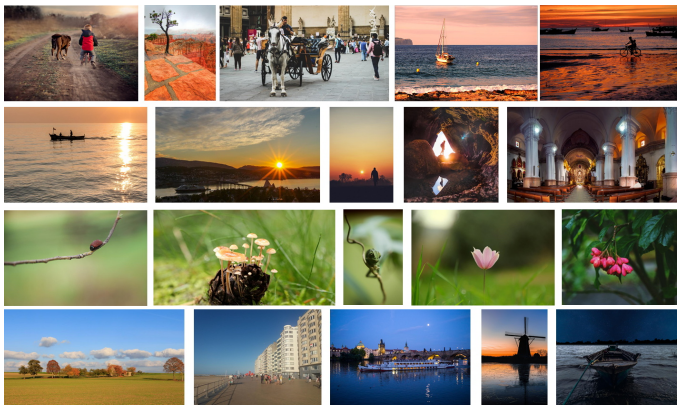


Fig. 5. Some sample images from the GAICD dataset.

where S_{crop} and S_{Image} represent the areas of crop and original image, respectively, and $\lambda \in [\frac{(M-2m+1)(N-2n+1)}{MN}, 1)$.

Aspect ratio: Because of the standard resolution of imaging sensors and displays, most people have been accustomed to the popular aspect ratios such as 16:9, 4:3 and 1:1. Candidate crops which have uncommon aspect ratios may be inconvenient to display and can make people feel uncomfortable. We thus require the aspect ratio of acceptable candidate crops satisfy the following condition:

$$\alpha_1 \leq \frac{W_{crop}}{H_{crop}} \leq \alpha_2, \tag{2}$$

where W_{crop} and H_{crop} are the width and height of a crop. Parameters α_1 and α_2 define the range of aspect ratio and we set them to 0.5 and 2 to cover most common aspect ratios.

With Eq. 1 and Eq. 2, the final number of candidate crops in each image is less than m^2n^2.

3.2 Database construction

Our proposed grid anchor based formulation reduces the number of candidate crops from $\frac{H(H-1)W(W-1)}{4}$ to less than m^2n^2. This enables us to annotate all the candidate crops for each image. To make the annotation cost as low as possible, we first made a small scale subjective study to find the smallest $\{M, N, m, n\}$ that ensure at least 3 acceptable crops for each image. We collected 100 natural images and invited five volunteers to participate in this study. We set $M = N \in \{16, 14, 12, 10\}$ and $m = n \in \{5, 4, 3\}$ to reduce possible combinations. λ in Eq. 1 was set to 0.5. After the tests, we found that $M = N = 12$ and $m = n = 4$ can lead to a good balance between cropping quality and annotation cost. Finally, the number of candidate crops is successfully reduced to no more than 90 for each image. Note that the setting of these parameters mainly aims to reduce annotation cost for training. In the testing stage, it is straightforward to use finer image grid to generate more candidate crops when necessary.

With the above settings, we constructed a Grid Anchor based Image Cropping Database (GAICD). We first crawled ~50,000 images from the Flickr website. Considering that many images uploaded to Flickr already have good composition, we manually selected 1,000 images whose composition can be obviously improved, as well as 236 images with proper composition to ensure the generality of the GAICD. The selected images have various aspect ratios and cover a variety of scenes and lighting conditions. There are 106,860 candidate crops of the 1,236 images in total. Some sample images from the GAICD are shown in Fig. 5.

Fig. 6. Interface of the developed annotation toolbox.

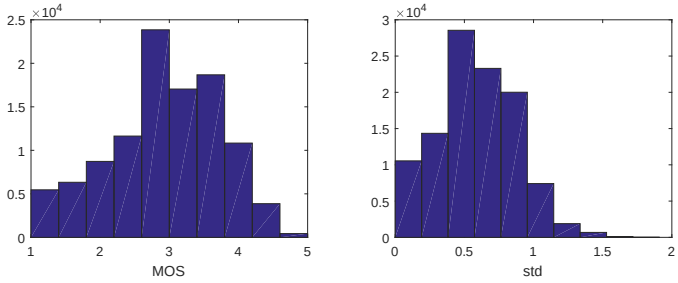


Fig. 7. Histograms of the MOS and standard deviation on the GAICD.

To improve the annotation efficiency, we developed an annotation toolbox whose interface is shown in Fig. 6. Each time, it displays one source image on the left side and 4 crops generated from it on the right side. The crops are displayed in ordered aspect ratio to alleviate the influence of dramatic changes of aspect ratio on human perception. Specifically, we choose six common aspect ratios (including 16:9, 3:2, 4:3, 1:1, 3:4 and 9:16) and group crops into six sets based on their closest aspect ratios. The top-right corner displays the approximate aspect ratio of current crops. Two horizontal and two vertical guidelines can be optionally used to assist judgement during the annotation. For each crop, we provide five scores (from 1 to 5, representing "bad," "poor," "fair," "good," and "excellent" crops) to rate by annotators. The annotators can either scroll their mouse or click the "Previous" or "Next" buttons to change page. In the bottom-left of the interface, we show the score distribution of rated crops for the current image as a reference for annotators. The bottom-right corner shows the progress of the annotation and the elapsed time.

A total of 19 annotators passed our test on photography composition and participated into the annotation. They are either experienced photographers from photography communities or students from the art department of two universities. Each crop was annotated by seven different subjects. The mean opinion score (MOS) was calculated for each candidate crop as its groundtruth quality score. The histograms of the mean opinion score (MOS) and standard deviation among the 106,860 candidate crops are plotted in Fig. 7. It can be seen that most crops have ordinary or poor quality, while about 10% crops have MOS larger than 4. Regarding to the standard deviation, only 5.75% crops are larger than 1, which indicates the consistency of annotations under our grid anchor based formulation. Fig. 8 shows one source image and several of its annotated crops (with MOS scores) in the GAICD.

Compared to the previous cropping datasets on which only one bounding box or several ranking pairs are annotated, our dataset has much more dense annotation and brings two significant benefits. First, our dense annotation provides not only richer but also finer supervised information for training cropping models. Second, the dense annotation enables us to define more reliable evaluation metrics on our new dataset, providing a more reasonable cropping benchmark for researchers to develop and evaluate their models.

3.3 Evaluation metrics

As shown in Table 1 and Fig. 1, the IoU and BDE metrics used in previous studies of image cropping are problematic. The dense annotations of our GAICD enable us to define more reliable metrics to evaluate cropping performance. Specifically, we define three types of metrics on our GAICD. The first tpye of metrics evaluate the ranking correlation between model's predictions and the groundtruth scores; the second type of metrics measure the model's performance to return the best crops; and the third type of metrics consider the ranking information into the best returns.

Ranking correlation metrics: The Pearson correlation coefficient (PCC) [41] and Spearman's rank-order correlation coefficient (SRCC) [42] can be naturally employed to evaluate the model's prediction consistency with the groundtruth MOS in our GAICD. These two metrics have been widely used in image quality and aesthetic assessment [43], [44], [45]. Denote by $\mathbf{g}_i$ the vector of MOS of all crops for image i, and by $\mathbf{p}_i$ the predicted scores of these crops by a model. The PCC is defined as:

$$PCC(\mathbf{g}_i, \mathbf{p}_i) = \frac{\text{cov}(\mathbf{g}_i, \mathbf{p}_i)}{\sigma_{\mathbf{g}_i}\sigma_{\mathbf{p}_i}}, \tag{3}$$

where cov and σ are the operators of the covariance and standard deviation. One can see that the PCC measures the linear correlation between two variables.

The SRCC is defined as the PCC between the rank variables:

$$SRCC(\mathbf{g}_i, \mathbf{p}_i) = \frac{\text{cov}(\mathbf{r}_{\mathbf{g}_i}, \mathbf{r}_{\mathbf{p}_i})}{\sigma_{\mathbf{r}_{\mathbf{g}_i}}\sigma_{\mathbf{r}_{\mathbf{p}_i}}}, \tag{4}$$

where $\mathbf{r}_{\mathbf{g}_i}$ and $\mathbf{r}_{\mathbf{p}_i}$ record the ranking order of scores in $\mathbf{g}_i$ and $\mathbf{p}_i$, respectively. The SRCC assesses the monotonic relationship between two variables. Given a testing set with T images, we calculate the average PCC and average SRCC over the T images as the final results:

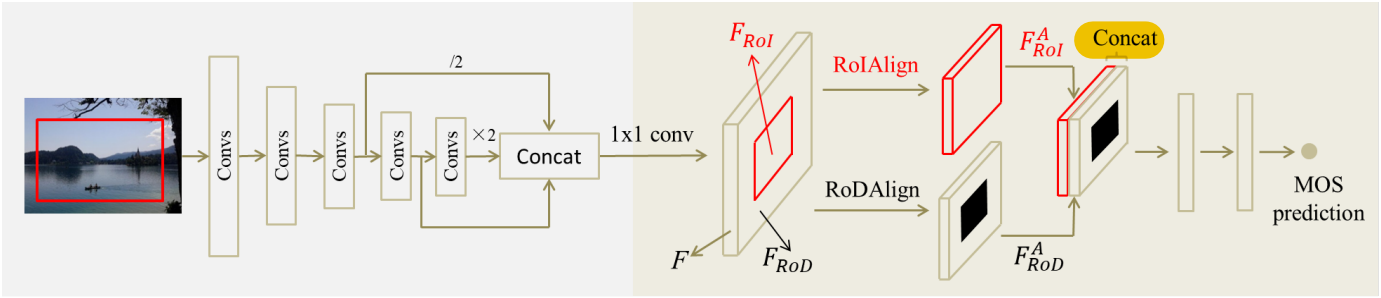
$$\overline{PCC} = \frac{1}{T}\sum_{i=1}^{T} PCC(\mathbf{g}_i, \mathbf{p}_i), \tag{5}$$

$$\overline{SRCC} = \frac{1}{T}\sum_{i=1}^{T} SRCC(\mathbf{g}_i, \mathbf{p}_i). \tag{6}$$

Best return metrics: Considering that in practical cropping applications, users care more about whether the cropping model can return the best crops rather than accurately rank all the candidate crops, we define a set of metrics to evaluate the models' ability to return the best crops. This new set of metrics is called as "return K of top-N" accuracy, which is similar to the "Precision at K" metric [46] widely used in modern retrieval systems. Specifically, we define the best crops of image i as the set of crops whose MOS rank top-N, and we denote this top-N set by $S_i(N)$. Suppose a cropping model returns K crops that have the highest prediction scores. We denote these K crops by $\{c_{ik}\}_{k=1}^{K}$ for image i. The "return K of top-N" accuracy checks on average how many of the returned K crops fall into the top-N set $S_i(N)$. It is defined as:

$$Acc_{K/N} = \frac{1}{TK}\sum_{i=1}^{T}\sum_{k=1}^{K} True(c_{ik} \in S_i(N)), \tag{7}$$

Fig. 8. One example source image and several of its annotated crops in our GAICD. The MOS is marked under each crop.



(a) Multi-scale feature extraction (b) Image cropping modeling

Fig. 9. The proposed CNN architecture for image cropping model learning. It consists of a multi-scale feature extraction module and a carefully designed cropping modeling module. Each convolutional block contains several convolution, batch normalization and ReLU layers. Symbols "×2" and "/2" represent bilinear upsampling and downsampling, respectively.

where $True(*) = 1$ if * is true, otherwise $True(*) = 0$. In practice, the number K of returned crops should not be set too large for users' convenience. In the ideal case, a cropping model should return only 1 crop to meet the user's expectation. For more patient users, no more than 4 crops could be returned to them. We thus set K to 1, 2, 3 and 4 in our benchmark. Regarding the selection of N, the statistic of MOS discussed in Section 3.2 shows that about 10% crops have MOS larger than 4, which means that there are on average 10 good crops for each image. We thus set N to 5 or 10. As a result, we obtain 8 accuracy indexes $Acc_{K/N}$ based on the different combinations of K and N.

Rank weighted best return metrics: The metric $Acc_{K/N}$ does not distinguish the rank among the returned top-N crops. For example, the value of $Acc_{1/5}$ will be the same when returning either the rank-1 or rank-5 crop. To further distinguish the rank of the returned top-N crops, we introduce a set of rank weighted best return metrics. Given the returned K crops of image i and their ranks among all the candidate crops, denoted by $\{r_{ik}\}_{k=1}^{K}$, we sort the K crops to have descending MOS, and obtain the sorted K crops $\{c_{ij}\}_{j=1}^{K}$ associated with their ranks $\{r_{ij}\}_{j=1}^{K}$. The "rank weighted return K of top-N" accuracy is defined as:

$$Acc^{w}_{K/N} = \frac{1}{TK}\sum_{i=1}^{T}\sum_{j=1}^{K} True(c_{ij} \in S_i(N)) * w_{ij}, \quad (8)$$

where

$$w_{ij} = e^{\frac{-\beta(r_{ij}-j)}{N}}, \quad (9)$$

where $\beta > 0$ is a scaling parameter and we simply set it to 1. The weight w_{ij} is designed under two considerations. First, w_{ij} should be larger if the crop c_{ij} has better rank. Second, w_{ij} should be 1 if the sorted rank r_{ij} matches the order of c_{ij} among the K returns, making the rank weighted accuracy $Acc^{w}_{K/N}$ able to reach 1 when the best crop set is returned.

We give an example to illustrate the calculation of $Acc^{w}_{4/5}$ for an input image. Suppose the returned 4 crops are ranked as $\{r_{ik}\}_{k=1}^{K} = \{2, 5, 3, 10\}$ among all candidate crops, it is easy to have $Acc_{4/5}$ = 0.75 since three are 3 crops falling into the top-5 set. The sorted ranks of the four returns are $\{r_{ij}\}_{j=1}^{K} = \{2, 3, 5, 10\}$, and the rank weighted accuracy is calculated as $Acc^{w}_{4/5} = \frac{1}{4}(e^{-\frac{2-1}{5}} + e^{-\frac{3-2}{5}} + e^{-\frac{5-3}{5}}) = 0.5769$. Compared with $Acc_{K/N}$, the metric $Acc^{w}_{K/N}$ can more precisely distinguish the quality of returns.

4 Cropping Model Learning

Limited by the insufficient amount of training data, most previous cropping methods focused on how to leverage additional aesthetic databases [11], [21], [22] or how to construct more training pairs [10], [24], paying limited attention to how to design a more suitable network for image cropping itself. They usually adopt the standard CNN architecture widely used in object detection. Our GAICD provides a better platform with much more annotated samples for model training. By considering the special properties of image cropping, we design an effective and lightweight cropping model. The overall architecture is shown in Fig. 9, which consists of a multi-scale feature extraction module and a carefully designed image cropping module. We also employ a set of data augmentation operations for learning robust cropping models.

4.1 Multi-scale feature extraction module

Efficient base model: A practical cropping model needs to be lightweight and efficient enough to be deployed on resource limited devices. Instead of employing those classical pre-trained CNN models such as AlexNet [47], VGG16 [29] or ResNet50 [30] as in previous work [10], [11], [13], [20], [21], [22], [23], [24], [28], we choose the more efficient architectures including the MobileNetV2

[26] and ShuffleNetV2 [27]. Fortunately, we found that using such efficient models will not sacrifice the cropping accuracy compared with their complicated counterparts, mostly owing to the special properties of image cropping and more advanced architecture designs of the MobileNetV2 and ShuffleNetV2. More details and discussions can be found in the ablation experiments.

Multi-scale features: As illustrated by the two examples in Fig. 10, the scale of objects varies significantly in different scenes. The features should also be responsive to the local distracting contents which should be removed in the final crop. As shown in Fig. 9(a), we extract multi-scale features from the same backbone CNN model. It has been widely acknowledged that the shallower CNN layers tend to capture the local textures while the deeper layers model the entire scene [48]. This motivates us to concatenate the feature maps from three different layers. Since the feature maps in different layers have different spatial resolution, we use bilinear downsampling and upsampling to make them have the same spatial resolution. The three feature maps are concatenated along the channel dimension as the output feature map.

4.2 Cropping module

Modeling both the RoI and RoD: One special property of image cropping is that we need to consider not only the region of interest (RoI) but also the region to be discarded (hereafter we call it region of discard (RoD)). On one hand, removing distracting information can significantly improve the composition. On the other hand, cutting out important region can dramatically change or even destroy an image. Taking the second last crop in Fig. 8 as an example, although it may have acceptable composition, its visual quality is much lower than the source image because the beautiful sunset glow is cropped out. The discarded information is unavailable to the cropping model if only the RoI is considered, while modeling the RoD can effectively solve this problem.

Referring to Fig. 9, denote by F the whole feature map output by the feature extraction module, and denote by F_{RoI} and F_{RoD} the feature maps in RoI and RoD, respectively. We first employ the RoIAlign operation [49] to transform F_{RoI} into F^A_{RoI}, which has fixed spatial resolution $s \times s$. The F_{RoD} is constructed by removing F_{RoI} from F, namely, setting the values of F_{RoI} to zeros in F. Then the RoDAlign operation (using the same bilinear interpolation as RoIAlign) is performed on F_{RoD}, resulting in F^A_{RoD} which has the same spatial resolution as F^A_{RoI}. F^A_{RoI} and F^A_{RoD} are concatenated along the channel dimension as one aligned feature map which contains the information of both RoI and RoD. The combined feature map is fed into two fully connected layers for final MOS prediction. Throughout our experiments, we fix s as 9 so that the bilinear interpolation in RoIAlign and RoDAlign can effectively leverage the entire feature map output by our feature extraction module. To be more specific, an input image of resolution 256×256 results in 16×16 feature maps after our feature extraction module, and bilinear interpolation takes four points to interpolate one point.

Modeling the spatial arrangement: The spatial arrangement of context and objects in an image plays a key role in image composition. For example, the most commonly used "rule of thirds" composition rule suggests to place important compositional elements at certain locations of an image [50]. Specifically, an image can be divided into nine parts by two equally spaced horizontal lines and two equally spaced vertical lines, and important elements should be placed along these lines or at the intersections of these lines, as shown in Fig 10. Other common composition rules such as symmetry and leading line also have certain spatial pattern. Considering that the downsampling and pooling operations after the feature extraction stage can cause significant loss of spatial information, we employ a fully-connected layer with large kernel size to explicitly model the spatial arrangement of an image. Our experimental results validate the advantage of this design in both cropping accuracy and efficiency than using several convolutional layers.

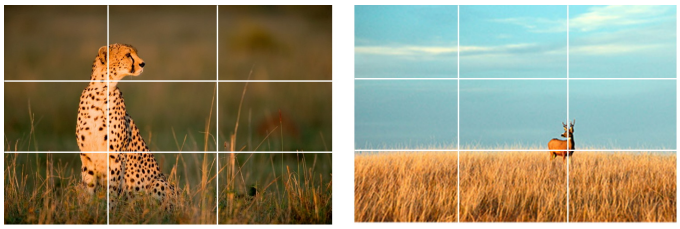


Fig. 10. Two examples using "rule of thirds" [50] composition.

Reducing the channel dimension: Another characteristic of image cropping is that it does not need to accurately recognize the category of different objects or scenes, which allows us to significantly reduce the channel dimension of the feature map. In practice, we found that the feature channel dimension can be reduced from several hundred to only 8 by using 1×1 convolution without sacrificing much the performance. The low channel dimension makes our image cropping module very efficient and lightweight.

Loss function: Denote by $e_{ij} = g_{ij} - p_{ij}$, where g_{ij} and p_{ij} are the groundtruth MOS and predicted score of the j-th crop for image i. The Huber loss [51] is employed as the loss function to learn our cropping model because of its robustness to outliers:

$$\mathcal{L}_{ij} = \begin{cases} \frac{1}{2}e_{ij}^2, \text{when } |e_{ij}| \leq \delta, \\ \delta|e_{ij}| - \frac{1}{2}\delta^2, \text{otherwise}, \end{cases} \tag{10}$$

where δ is fixed at 1 throughout our experiments.

4.3 Data augmentation

Data augmentation is an effective way to improve the robustness and performance of deep CNN models. However, many popular data augmentation operations are inappropriate for cropping. For example, rotation and vertical flipping can severely destroy the composition. Since the IoU is unreliable for evaluating cropping performance, randomly generating crops and assigning labels to them based on IoU [21] is also questionable. We thus employ a set of operations, which do not affect the composition, for data augmentation. Specifically, we randomly adjust the brightness, contrast, saturation, hue and horizontally flip the input image in the training stage.

5 Experiments

5.1 Implementation details

We randomly selected 200 images from our GAICD as the testing set and used the remaining 1,036 images (containing 89,519 annotated crops in total) for training and validation. In the training stage, our model takes one image and 64 randomly selected crops of it as a batch to input. In the testing stage, the trained

TABLE 2
Image cropping performance by using different feature extraction modules. The FLOPs are calculated on image with 256×256 pixels. All the single-scale models use the feature map after the fourth convolutional block which has the best performance among the three scales. The last convolutional block in both the MobileNetV2 and ShuffleNetV2 contains most of the parameters because of the high channel dimension, while it is simply a max pooling layer in VGG16 model and does not have any parameter.

Base model	Scale	Aug.	FLOPs	# of params	$\overline{SRCC}$	$\overline{PCC}$	$Acc_{1/5}$	$Acc_{4/5}$	$Acc_{1/10}$	$Acc_{4/10}$	$Acc^{w}_{4/5}$	$Acc^{w}_{4/10}$
VGG16	Single	No	22.3G	14.7M	0.752	0.778	58.0	47.7	74.0	67.9	32.2	49.2
	Single	Yes	22.3G	14.7M	0.764	0.791	59.5	49.2	76.0	69.3	33.3	50.3
	Multi	Yes	22.3G	14.7M	**0.777**	**0.800**	**60.5**	**50.2**	**77.5**	**70.6**	**34.4**	**51.3**
MobileNetV2	Single	No	314M	0.54M	0.760	0.782	58.5	49.1	75.5	69.0	33.6	51.6
	Single	Yes	314M	0.54M	0.775	0.793	60.5	51.4	77.5	70.9	35.1	53.2
	Multi	Yes	407M	1.81M	**0.783**	**0.806**	**62.5**	**52.5**	**78.5**	**72.3**	**36.2**	**54.4**
ShuffleNetV2	Single	No	126M	0.28M	0.751	0.780	58.0	48.8	76.0	68.1	34.2	51.1
	Single	Yes	126M	0.28M	0.763	0.792	60.0	50.9	77.5	70.3	35.8	52.6
	Multi	Yes	170M	0.78M	**0.774**	**0.801**	**61.5**	**52.0**	**78.5**	**71.3**	**37.2**	**53.6**

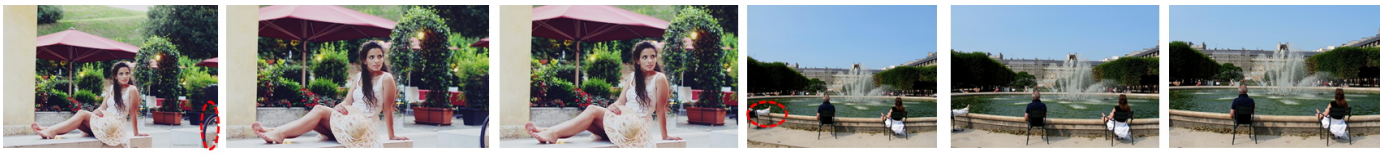


Fig. 11. Qualitative comparison between single-scale and multi-scale feature based crops. Using multi-scale features can effectively detect and remove local distracting elements that tend to be ignored by single-scale feature.

Fig. 12. Qualitative comparison of modeling only RoI against modeling both RoI and RoD. Modeling both RoI and RoD can preserve as much useful information as possible in the source image, while modeling only the RoI may lead to unnecessary information loss.

model evaluates all the generated crops of one image and outputs a predicted MOS for each crop. To improve the training and testing efficiency, the short side of input images is resized to 256. The feature extraction module employs the CNN models pre-trained on the ImageNet dataset. The cropping modeling module is randomly initialized using the method proposed in [52]. The standard ADAM [53] optimizer with the default parameters is employed to train our model for 80 epoches. Learning rate is fixed at $1e^{-4}$ throughout our experiments. The RGB values in input images are scaled to the range of [0,1] and normalized using the mean and standard deviation calculated on the ImageNet. The MOS are normalized by removing the mean and dividing by the standard deviation across the training set. More implementation details can be found in our released code.

5.2 Ablation study

5.2.1 Feature extraction module

We first conduct a set of ablation studies to evaluate the performance of different base models for feature extraction, single-scale and multi-scale features and data augmentation for model training. The three different base models include VGG16 [29], MobileNetV2 [26] and ShuffleNetV2 [27]. The width multiplier is set to 1.0 for both MobileNetV2 and ShuffleNetV2. In these experiments, the image cropping module (including both the RoI and RoD) is fixed for all cases except that the 1×1 convolutional layer for dimension reduction has different input dimension for different models. Since the accuracy indexes have similar tendency, we only report several representative indexes for each type of metrics in the ablation study to save space. The specific setting, model complexity and cropping performance for each case are reported in Table 2.

Base model: We found that the lightweight MobileNetV2 and ShuffleNetV2 obtain even better performance than the VGG16 model on all the three types of evaluation metrics. This is owning to the more advanced architecture designs of MobileNetV2 and ShuffleNetV2, both of which leverage many latest useful practices in CNN architecture design such as residual learning, batch normalization and group convolution. It is worth mentioning that their classification accuracies are also comparable or slightly better than the VGG16 model on the ImageNet dataset. Besides, the lightweight networks of MobileNetV2 and ShuffleNetV2 are easier to be trained than VGG16 considering the fact that our GAICD is still not very big. Between MobileNetV2 and ShuffleNetV2, the former obtains slightly better performance (at the same width multiplier) on most of the metrics, which is consistent to their relative performances in other vision tasks [27]. Regarding the computational cost, both the number of parameters and FLOPs (the number of multiply-adds) of MobileNetV2 and ShuffleNetV2 are more than one order smaller the the VGG16.

Multi-scale features: As can been seen from Table 2, ex-

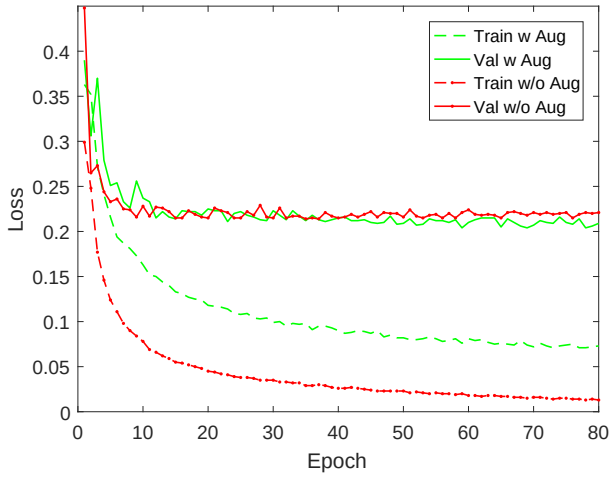


Fig. 13. Learning curves with and without data augmentation by the ShuffleNetV2.

tracting multi-scale features improves the performance for all the three base models. As we have mentioned before, we extract three scales of features from the same backbone network. The single-scale models employed in this study only used the feature map after the fourth convolutional block which was found to have the best performance among the three scales. Regarding the computational cost, extracting multi-scale features only needs to calculate one additional (i.e. the fifth) convolutional block, and the FLOPs only increase by about $\frac{1}{3}$ compared with the single-scale counterparts. Thus the multi-scale models are still very lightweight and efficient. A qualitative comparison of the cropping results by single- and multi-scale features on two images is shown in Fig. 11. The cropping results show that using multi-scales features can effectively remove local distracting elements that may be ignored by single-scale features.

Data augmentation: The results in Table 2 show that data augmentation consistently improves the performance in terms of all the employed metrics for all the three base models. The learning curves with and without data augmentation by ShuffleNetV2 are plotted in Fig. 13, where we randomly selected 100 images from the training set for validation. One can see that, without data augmentation, the loss decreases very fast on the training set but it has a significant gap to the loss on the validation set. By using data augmentation, smaller loss can be obtained on the validation set, improving the generalization capability of the trained model.

5.2.2 Image cropping module

We then evaluate the three special designs in the proposed image cropping module, including RoI and RoD modeling, large kernel size and low channel dimension.

RoI and RoD: We evaluate the roles of RoI and RoD on both MobileNetV2 and ShuffleNetV2 with all the other settings fixed. The results of modeling only RoI, only RoD and both of them are reported in Table 3. As can be seen, modeling only the RoD obtains unsatisfied performance, modeling only the RoI performs much better, while modeling simultaneously the RoI and RoD achieves the best cropping accuracy in all cases. A qualitative comparison of modeling only RoI against modeling both RoI and RoD is shown in Fig. 12. One can observe that modeling both RoI and RoD can preserve as much useful information as possible in the source image while modeling only the RoI may lead to some information loss. This corroborates our analysis that image cropping needs to consider both the RoI and RoD.

Kernel size: Given the feature map after RoIAlign and RoDAlign with 9×9 spatial resolution, we propose to use one fully-connected (FC) layer with a large kernel to explicitly modeling the spatial arrangement of the feature map rather than using several small size stride convolutional (Conv) layers. A comparison of using one single $9 \times 9 \times 16 \times 768$ FC layer, two $5 \times 5 \times 16 \times 768$ Conv layers (the first layer uses stide 2, followed by a $1 \times 1 \times 768 \times 16$ Conv layer for dimension reduction) and three $3 \times 3 \times 16 \times 768$ Conv layers (the first two layers use stide 2, each followed by a $1 \times 1 \times 768 \times 16$ Conv layer) are listed in the top half of Table 4. One can see that using one single FC layer obtains higher performance than its competitors. This is because the spatial information may be lost in the downsampling process by stride convolution. Regarding the computational cost, using one single FC layer also has smaller FLOPs since each element calculates only once, while the feature map is repeatedly calculated in the other two cases.

Channel dimension reduction: We also evaluate a set of channel dimensions for the FC layer and report the results in the bottom half of Table 4. Given the multi-scale feature map output by ShuffleNetV2 with 812 channels, we can reduce the channel dimension to only 8 with little performance decay. Note that the channel dimension in the kernel is double of the that in the feature map because of concatenation of the RoI and RoD branches. The performance is still reasonable even if we reduce the channel dimension to 1. Benefiting from the low channel dimension, the FLOPs of our cropping modeling module is only 1.0M, which is almost ignorable compared to the FLOPs in feature extraction.

5.3 Comparison to other methods

5.3.1 Comparison methods

Though a number of image cropping methods have been developed [10], [11], [13], [20], [21], [22], [23], [24], many of them do not release the source code or executable program. We thus compare our method, namely Grid Anchor based Image Cropping (GAIC), with the following baseline and recently developed state-of-the-art methods whose source codes are available.

Baseline_L: The baseline_L does not need any training. It simply outputs the largest crop among all eligible candidates. The result is similar to the "baseline_N" mentioned in Table 1, i.e., the source image without cropping.

VFN [21]: The View Finding Network (VFN) is trained in a pair-wise ranking manner using professional photographs crawled from the Flickr. High-quality photos were first manually selected, and a set of crops were then generated from each image. The ranking pairs were constructed by always assuming that the source image has better quality than the generated crops.

VEN and VPN [24]: Compared with VFN, the View Evaluation Network (VEN) employs more reliable ranking pairs to train the model. Specifically, the authors annotated more than 1 million ranking pairs using a two-stage annotation strategy. A more efficient View Proposal Network (VPN) was proposed in the same work, and it was trained using the predictions of VEN. The VPN is based on the detection model SSD [40], and it outputs a prediction vector for 895 predefined boxes.

A2-RL [13]: The A2RL is trained in an iterative optimization manner. The model adjusts the cropping window and calculates a reward (based on predicted aesthetic score) for each step.

TABLE 3
Ablation experiments on the roles of RoI and RoD.

Base model	module	$\overline{SRCC}$	$\overline{PCC}$	$Acc_{1/5}$	$Acc_{4/5}$	$Acc_{1/10}$	$Acc_{4/10}$	$Acc^{w}_{1/5}$	$Acc^{w}_{4/5}$	$Acc^{w}_{1/10}$	$Acc^{w}_{4/10}$
MobileNetV2	RoD	0.672	0.715	45.0	39.8	61.0	56.6	31.9	26.4	43.2	41.3
	RoI	0.770	0.792	60.5	51.4	76.5	71.1	37.1	34.6	55.3	52.4
	RoI+RoD	**0.783**	**0.806**	**62.5**	**52.5**	**78.5**	**72.3**	**39.6**	**36.2**	**56.9**	**54.4**
ShuffleNetV2	RoD	0.678	0.718	45.0	39.1	61.5	55.7	32.4	28.0	44.6	41.7
	RoI	0.764	0.785	59.5	50.1	76.5	69.6	39.2	35.4	55.4	51.6
	RoI+RoD	**0.774**	**0.801**	**61.5**	**52.0**	**78.5**	**71.3**	**40.3**	**37.2**	**57.3**	**53.6**

TABLE 4
Image cropping performance by using different number and size of kernels in the cropping modeling module. The ShuffleNetV2 model is employed as the feature extraction module for all cases. Note that the channel dimension in the kernel is double of that in the feature map because of concatenation of the RoI and RoD branches.

kernels	FLOPs	$\overline{SRCC}$	$\overline{PCC}$	$Acc_{1/5}$	$Acc_{4/5}$	$Acc_{1/10}$	$Acc_{4/10}$	$Acc^{w}_{1/5}$	$Acc^{w}_{4/5}$	$Acc^{w}_{1/10}$	$Acc^{w}_{4/10}$
$\left[\begin{smallmatrix}3,3,16,768\end{smallmatrix}\right]\times 3$	4.28M	0.765	0.785	57.5	48.6	74.0	68.8	37.1	33.4	53.0	49.8
$\left[\begin{smallmatrix}5,5,16,768\end{smallmatrix}\right]\times 2$	8.28M	0.769	0.795	60.5	51.2	76.5	70.1	38.7	35.1	55.9	52.4
$\left[\begin{smallmatrix}9,9,16,768\end{smallmatrix}\right]\times 1$	1.00M	**0.774**	**0.801**	**61.5**	**52.0**	**78.5**	**71.3**	**40.3**	**37.2**	**57.3**	**53.6**
$\left[\begin{smallmatrix}9,9,64,768\end{smallmatrix}\right]\times 1$	3.98M	**0.780**	**0.806**	**62.5**	**52.5**	79.0	**71.8**	40.5	37.4	**57.7**	**54.1**
$\left[\begin{smallmatrix}9,9,32,768\end{smallmatrix}\right]\times 1$	1.99M	0.777	0.804	62.0	52.2	**79.5**	71.5	**40.7**	**37.5**	57.5	53.8
$\left[\begin{smallmatrix}9,9,16,768\end{smallmatrix}\right]\times 1$	1.00M	0.774	0.801	61.5	52.0	78.5	71.3	40.3	37.2	57.3	53.6
$\left[\begin{smallmatrix}9,9,8,768\end{smallmatrix}\right]\times 1$	0.50M	0.767	0.793	62.0	51.6	78.0	70.7	39.5	36.6	56.8	53.1
$\left[\begin{smallmatrix}9,9,4,768\end{smallmatrix}\right]\times 1$	0.25M	0.760	0.785	61.0	50.7	77.0	69.5	38.6	35.5	56.1	52.1
$\left[\begin{smallmatrix}9,9,2,768\end{smallmatrix}\right]\times 1$	0.13M	0.752	0.775	59.0	48.4	75.0	67.5	37.1	34.1	54.8	50.3

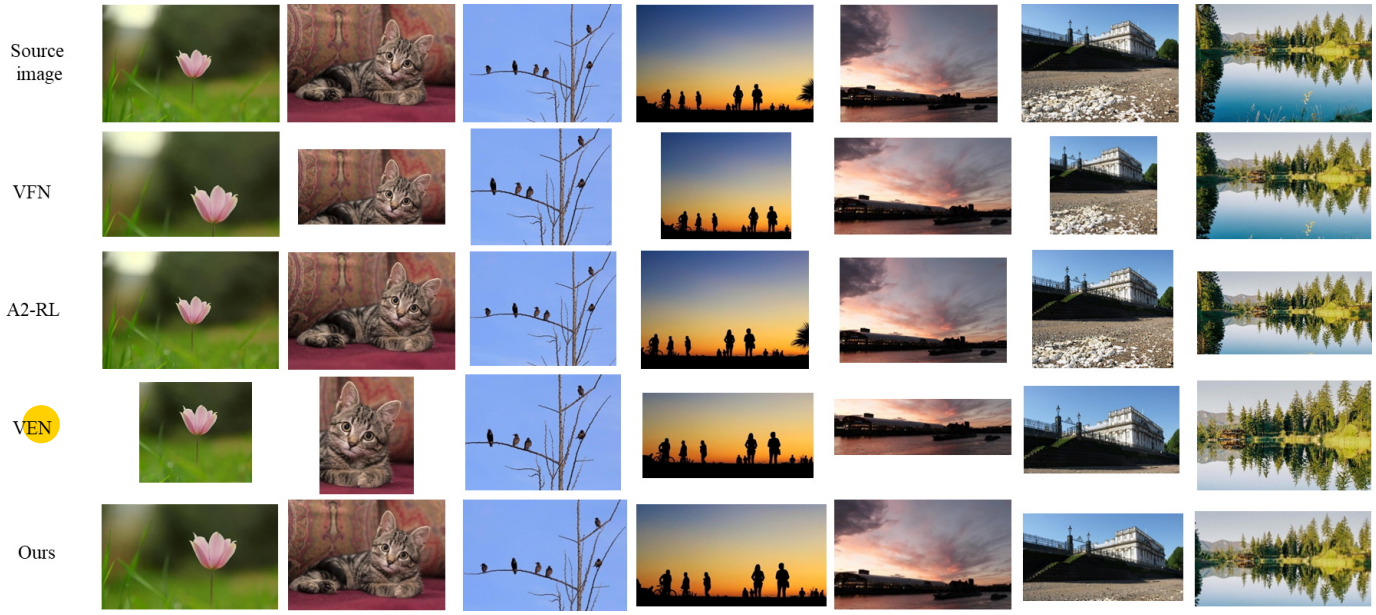


Fig. 14. Qualitative comparison of returned top-1 crops by different methods.

The iteration stops when the accumulated reward satisfies some termination criteria.

5.3.2 Qualitative comparison

To demonstrate the advantages of our cropping method over previous ones, we first conduct qualitative comparison of different methods on various scenes including single object, multi-objects, building and landscape. Note that these images are out of any existing cropping databases. In the first set of comparison, we compare all methods under the setting of returning only one best crop. Each model uses its default candidate crops generated by its source code except for VFN, which does not provide such code and uses the same candidates as our method. The results are shown in Fig. 14. We can make several interesting observations. Both VFN and A2-RL fail to robustly remove distracting elements in images. VFN sometimes cuts out important content, while A2-RL simply returns the source image in many cases. VEN and our GAIC model can stably output visually pleasing crops. The major differences lie in that VEN prefers more close-up crops while our GAIC tends to preserve as much useful information as possible.

A flexible cropping system should be able to output acceptable results under different requirements in practice, e.g., different aspect ratios. In this case, we generate multi-scale candidate crops with fixed aspect ratio and feed the same candidate crops into each

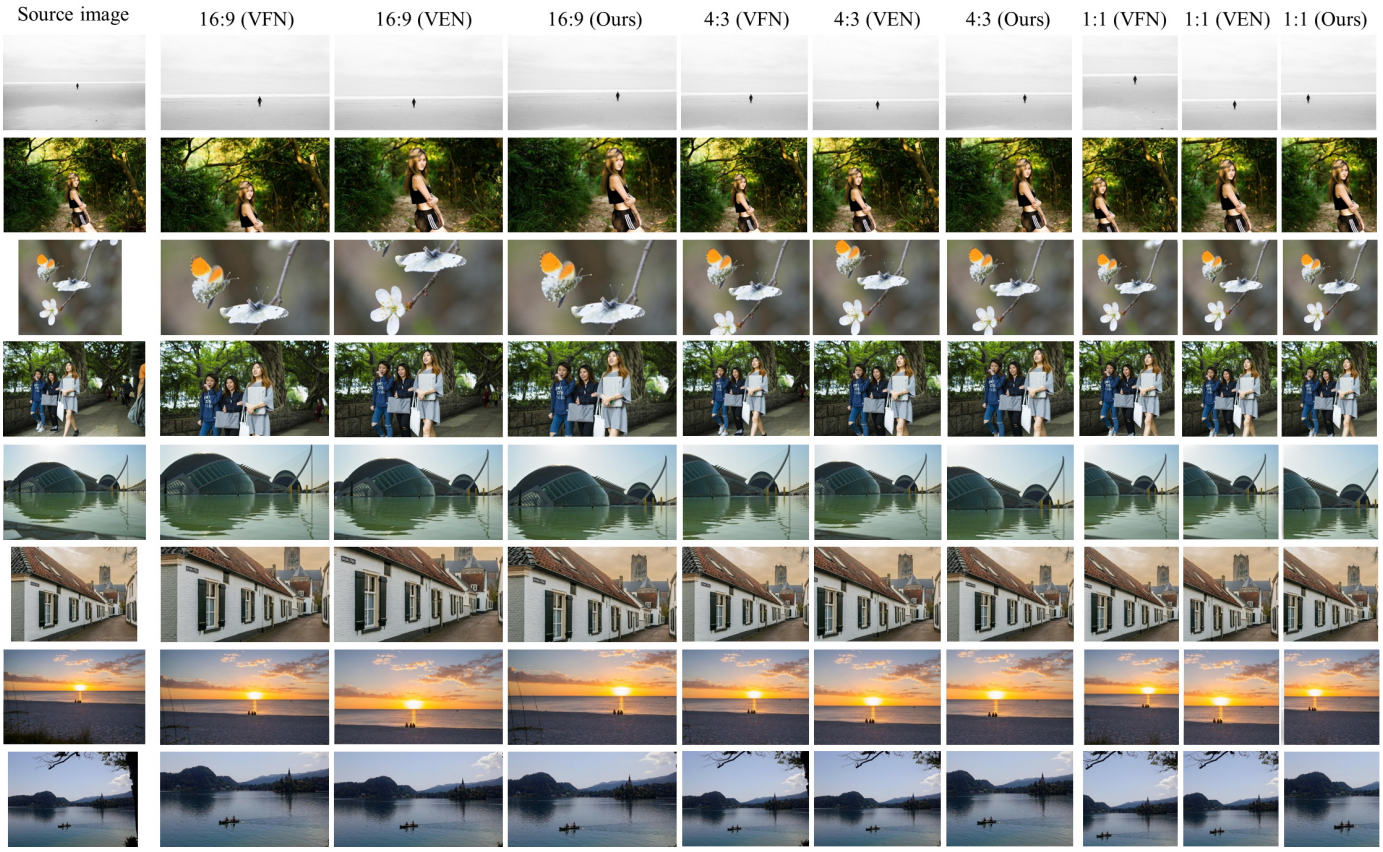


Fig. 15. Qualitative comparison of returning crops with different aspect ratios by different methods.

of the competing models. In Fig. 15, we show the top-1 returned crops by the competing methods under three most commonly used aspect ratios: 16:9, 4:3 and 1:1. The A2-RL is not included because it does not support this test. Again, our model outputs the most visually pleasing crops in most cases.

5.3.3 Quantitative comparison

We then perform quantitative comparisons by using the metrics defined in Section 3.3. Among the competitors, VFN, VEN and our GAIC support predicting scores for all the candidate crops provided by our database, thus they can be quantitatively evaluated by all the defined evaluation metrics. VPN uses its own pre-defined cropping boxes which are different from our database, and Baseline_L and A2-RL output only one single crop. Therefore, we can only calculate $Acc_{1/5}$, $Acc^{w}_{1/5}$, $Acc_{1/10}$ and $Acc^{w}_{1/10}$ for them. We approximate the output boxes by VPN and A2-RL to the nearest anchor box in our database when calculating these accuracy indexes. The results of all competing methods on all the defined metrics are shown in Table 5.

We can draw several conclusions from the quantitative results. First, one can see that both A2-RL and VFN only obtain comparable performance to Baseline_L. This is mainly because A2-RL is supervised by a general aesthetic classifier in training while the general aesthetic supervision across images cannot accurately discriminate different crops within one image, and the ranking pairs used in VFN are not very reliable because crops generated by well-composed images do not necessarily have worse composition than the source image. Although using the same pairwise learning strategy, VEN obtains much better performance than VFN by collecting more reliable ranking pairs through human annotations, which proves the necessity of human annotations for the cropping task. VPN performs slightly worse than VEN as expected because it is supervised by the predictions of VEN. Our model in the conference version already outperforms VEN by a large margin on all the evaluation metrics, benefitting from our dense annotated dataset which provides richer supervised information compared to the pair-wise ranking annotations used by VEN. Employing more efficient CNN architectures and more effective multi-scale features, our new models further significantly boost the cropping performance than our conference version on all the metrics.

5.3.4 Running speed comparison

A practical image cropping model should also have fast speed for real-time implementation. In the last two columns of Table 5, we compare the running speed in terms of frame-per-second (FPS) on both GPU and CPU for all competing methods. All models are tested on the same PC with i7-6800K CPU, 64G RAM and one GTX 1080Ti GPU, and our method is implemented under the PyTorch toolbox. As can be seen, our GAIC model based on the MobileNetV2 runs at 200 FPS on GPU and 6 FPS on CPU, and its counterpart based on the ShuffleNetV2 runs at 142 FPS on GPU and 12 FPS on CPU, both of which are much faster than the other competitors. It is worth mentioning that the GPU speeds in our testing are inconsistent with the CPU speeds because some operations such as group convolution and channel shuffle in the MobileNetV2 and ShuffleNetV2 are not well supported in PyTorch to make full use of the GPU computational capability. The other models are much slower because they either

TABLE 5
Quantitative comparison between different methods on the GAICD. "–" means that the result is not available. The reported FPS are tested on our own devices using one GTX 1080Ti GPU and i7-6800K CPU. [⋆] The GPU speeds are inconsistent with the CPU speeds because some operations such as group convolution and channel shuffle in the MobileNetV2 and ShuffleNetV2 are not well supported in PyTorch to make full use of the GPU computational capability.

Method	$Acc_{1/5}$	$Acc_{2/5}$	$Acc_{3/5}$	$Acc_{4/5}$	$Acc_{1/10}$	$Acc_{2/10}$	$Acc_{3/10}$	$Acc_{4/10}$	$\overline{SRCC}$	$\overline{PCC}$
Baseline_L	24.5	–	–	–	41.0	–	–	–	–	–
A2-RL [13]	23.0	–	–	–	38.5	–	–	–	–	–
VPN [24]	40.0	–	–	–	49.5	–	–	–	–	–
VFN [21]	27.0	28.0	27.2	24.6	39.0	39.3	39.0	37.3	0.450	0.470
VEN [24]	40.5	36.5	36.7	36.8	54.0	51.0	50.4	48.4	0.621	0.653
GAIC (Conf.) [28]	53.5	51.5	49.3	46.6	71.5	70.0	67.0	65.5	0.735	0.762
GAIC (Mobile-V2)	**62.5**	**58.3**	**55.3**	**52.5**	**78.5**	**76.2**	**74.8**	**72.3**	**0.783**	**0.806**
GAIC (Shuffle-V2)	61.5	56.8	54.8	52.0	**78.5**	75.5	73.8	71.3	0.774	0.801

Method	$Acc^w_{1/5}$	$Acc^w_{2/5}$	$Acc^w_{3/5}$	$Acc^w_{4/5}$	$Acc^w_{1/10}$	$Acc^w_{2/10}$	$Acc^w_{3/10}$	$Acc^w_{4/10}$	FPS (GPU)⋆	FPS (CPU)
Baseline_L	15.6	–	–	–	26.9	–	–	–	–	–
A2-RL [13]	15.3	–	–	–	25.6	–	–	–	5	0.05
VPN [24]	19.5	–	–	–	29.0	–	–	–	75	0.8
VFN [21]	16.8	13.6	12.5	11.1	25.9	22.1	20.7	19.1	0.5	0.005
VEN [24]	20.0	16.1	14.2	12.8	30.0	25.9	24.2	23.8	0.2	0.002
GAIC (Conf.) [28]	37.6	33.9	31.5	30.0	53.7	49.4	48.4	46.9	125	1.2
GAIC (Mobile-V2)	39.6	39.1	38.3	36.2	56.9	**56.5**	**55.9**	**54.4**	**200**	6
GAIC (Shuffle-V2)	**40.3**	**39.4**	**38.6**	**37.2**	**57.3**	55.0	54.7	53.6	142	**12**

Fig. 16. 16:9 crops generated by our model on images taken by wide lens action cameras. First row: 4:3 raw images captured by action cameras. Second row: 16:9 images generated by our model.

employ heavy CNN architectures (VPN, GAIC (Conf.)), or need to individually process each crop (VFN and VEN) or need to iteratively update the cropping window several times (A2-RL), making them hard to be used in practical applications with real-time implementation requirement.

5.3.5 *Results on previous datasets*

As discussed in the introduction section, the limitations of previous image cropping databases and evaluation metrics make them unable to reliably reflect the cropping performance of a method. Nonetheless, we still evaluated our model on the ICDB [5] and FCDB [10] using the IoU as metric for reference of interested readers. Since some groundtruth crops on these two databases have uncommon aspect ratios, we did not employ the aspect ratio constraint when generating candidate crops on these two datasets. We found that the value of λ defined in the area constraint (Eq. 1) largely affects the performance of our model on the ICDB. We tuned λ for the MobileNetV2 and ShuffleNetV2 based models and report their best results in Table 1. However, like most previous methods, our models still obtain even smaller IoU than the baselines on the ICDB dataset and slightly better result on the FCDB dataset. In contrast, as shown in previous subsections, a well trained model on our GAICD can obtain much better performance than the baseline. These results further prove the advantages of our new database as well as the associated metrics compared to previous ones.

5.4 Application to action cameras

We also evaluate the generalization capability of our model on a practical application: automatically cropping the images captured by action cameras. The action cameras usually have wide lens for capturing large field of view which is inevitably associated with severe lens distortion. We tested our trained model on 4:3 images taken by GoPro Hero 7 and DJI Osmo Pocket, and generated 16:9 crops. The results on six scenes are shown in Fig. 16. We found that the model trained on our dataset can generalize well to the images with large lens distortion, because the lens distortion does not severely change the spatial arrangement of image content.

6 Conclusion and Discussion

We analyzed the limitations of existing formulation and databases on image cropping, and proposed a more reliable and efficient formulation for practical image cropping, namely grid anchor based image cropping (GAIC). A new benchmark was constructed, which contains 1,236 source images and 106,860 annotated crops. Three new types of metrics were defined to reliably and comprehensively evaluate the cropping performance on our database. We also designed very lightweight and effective cropping models by

considering the special properties of cropping. Our GAIC model can robustly output visually pleasing crops under different aspect ratios. It runs at a speed up to 200 FPS on one GTX1080Ti GPU and 12 FPS on CPU, enabling real-time implementations on mobile devices.

There remain some limitations in our work, which leave much space for improvement. Firstly, our GAIC dataset is still limited in size considering the billions of photos generated in each day. It is expected that larger scale cropping dataset with reliable annotations can be constructed in the future. Second, the accuracies especially the rank weighted accuracies need further improvement. The cropping models are expected to learn more discriminative representations of photo composition in order to more accurately return the best crops.

REFERENCES

[1] Wikipedia contributors, "Cropping (image) — Wikipedia, the free encyclopedia," https://en.wikipedia.org/w/index.php?title=Cropping_(image)&oldid=847382681, 2018, [Online; accessed 10-July-2018]. 1

[2] L.-Q. Chen, X. Xie, X. Fan, W.-Y. Ma, H.-J. Zhang, and H.-Q. Zhou, "A visual attention model for adapting images on small displays," *Multimedia systems*, vol. 9, no. 4, pp. 353–364, 2003. 1, 3

[3] A. Chor, J. Schwartz, P. Hellyar, T. Kasperkiewicz, and D. Parlin, "System for automatic image cropping based on image saliency," Apr. 6 2006, US Patent App. 10/956,628. 1

[4] N. Jogo, "Image cropping and synthesizing method, and imaging apparatus," Apr. 24 2007, US Patent 7,209,149. 1

[5] J. Yan, S. Lin, S. Bing Kang, and X. Tang, "Learning the change for automatic image cropping," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 971–978. 1, 2, 3, 12

[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, pp. 1105–1108. 1, 2, 3, 4

[7] E. O. Downing, O. M. Koenders, and B. T. Grover, "Automated image cropping to include particular subjects," Apr. 28 2015, US Patent 9,020,298. 1

[8] N. Bhatt and T. Cherna, "Multifunctional environment for image cropping," Oct. 13 2015, US Patent 9,158,455. 1

[9] J. Chen, G. Bai, S. Liang, and Z. Li, "Automatic image cropping: A computational complexity study," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 507–515. 1

[10] Y.-L. Chen, T.-W. Huang, K.-H. Chang, Y.-C. Tsai, H.-T. Chen, and B.-Y. Chen, "Quantitative analysis of automatic image cropping algorithms: A dataset and comparative study," in *Proceeding of the IEEE Winter Conference on Applications of Computer Vision*, 2017, pp. 226–234. 1, 2, 3, 6, 9, 12

[11] W. Wang and J. Shen, "Deep cropping via attention box prediction and aesthetics assessment," in *Proceeding of the IEEE International Conference on Computer Vision*, 2017. 1, 2, 3, 6, 9

[12] C. S. B. Chedeau, "Image cropping according to points of interest," Mar. 28 2017, US Patent 9,607,235. 1

[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. 1, 2, 3, 6, 9, 12

[14] G. Ciocca, C. Cusano, F. Gasparini, and R. Schettini, "Self-adaptive image cropping for small displays," *IEEE Transactions on Consumer Electronics*, vol. 53, no. 4, 2007. 1

[15] B. Suh, H. Ling, B. B. Bederson, and D. W. Jacobs, "Automatic thumbnail cropping and its effectiveness," in *ACM symposium on User Interface Software and Technology*, 2003, pp. 95–104. 1, 3

[16] L. Marchesotti, C. Cifarelli, and G. Csurka, "A framework for visual saliency detection with applications to image thumbnailing," in *Proceeding of the IEEE International Conference on Computer Vision*, 2009, pp. 2232–2239. 1, 3

[17] A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. Cohen, "Gaze-based interaction for semi-automatic photo cropping," in *ACM SIGCHI*, 2006, pp. 771–780. 1, 3

[18] F. Stentiford, "Attention based auto image cropping," in *ICVS Workshop on Computation Attention & Applications*, 2007. 1, 3

[19] J. Freixenet, X. Muñoz, D. Raba, J. Martí, and X. Cufí, "Yet another survey on image segmentation: Region and boundary information integration," in *Proceedings of the European Conference on Computer Vision*, 2002, pp. 408–422. 1

[20] 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. 1, 3, 6, 9

[21] 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, pp. 37–45. 1, 2, 3, 6, 7, 9, 12

[22] Y. Deng, C. C. Loy, and X. Tang, "Aesthetic-driven image enhancement by adversarial learning," *arXiv preprint arXiv:1707.05251*, 2017. 1, 3, 6, 9

[23] G. Guo, H. Wang, C. Shen, Y. Yan, and H.-Y. M. Liao, "Automatic image cropping for visual aesthetic enhancement using deep neural networks and cascaded regression," *arXiv preprint arXiv:1712.09048*, 2017. 1, 3, 6, 9

[24] 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 the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5437–5446. 1, 2, 3, 6, 9, 12

[25] 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. 2, 3

[26] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520. 2, 7, 8

[27] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 116–131. 2, 7, 8

[28] H. Zeng, L. Li, Z. Cao, and L. Zhang, "Reliable and efficient image cropping: A grid anchor based approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2, 6, 12

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 2, 6, 8

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778. 2, 6

[31] 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. 3

[32] M. Zhang, L. Zhang, Y. Sun, L. Feng, and W. Ma, "Auto cropping for digital photographs," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2005. 3

[33] M. Nishiyama, T. Okabe, Y. Sato, and I. Sato, "Sensation-based photo cropping," in *ACM Multimedia*, 2009, pp. 669–672. 3

[34] B. Cheng, B. Ni, S. Yan, and Q. Tian, "Learning to photograph," in *ACM Multimedia*, 2010, pp. 291–300. 3

[35] L. Liu, R. Chen, L. Wolf, and D. Cohen-Or, "Optimizing photo composition," in *Computer Graphics Forum*, vol. 29, no. 2, 2010, pp. 469–478. 3

[36] 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. 3

[37] 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. 3

[38] N. Murray, L. Marchesotti, and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2408–2415. 3

[39] W. Luo, X. Wang, and X. Tang, "Content-based photo quality assessment," in *Proceeding of the IEEE International Conference on Computer Vision*, 2011, pp. 2206–2213. 3

[40] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Proceedings of the European Conference on Computer Vision*. Springer, 2016, pp. 21–37. 3, 9

[41] Wikipedia contributors, "Pearson correlation coefficient — Wikipedia, the free encyclopedia," https://en.wikipedia.org/w/index.php?title=Pearson_correlation_coefficient&oldid=905965350, 2019, [Online; accessed 16-July-2019]. 5

[42] Wikipedia contributors, "Spearman's rank correlation coefficient — Wikipedia, the free encyclopedia," https://en.wikipedia.org/w/index.php?

title=Spearman%27s_rank_correlation_coefficient&oldid=899964572, 2019, [Online; accessed 16-July-2019]. 5

[43] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *Proceedings of the European Conference on Computer Vision*. Springer, 2016, pp. 662–679. 5

[44] K. Ma, Z. Duanmu, Q. Wu, Z. Wang, H. Yong, H. Li, and L. Zhang, "Waterloo exploration database: New challenges for image quality assessment models," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 1004–1016, 2017. 5

[45] H. Talebi and P. Milanfar, "NIMA: Neural image assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3998–4011, 2018. 5

[46] Wikipedia contributors, "Evaluation measures (information retrieval) — Wikipedia, the free encyclopedia," https://en.wikipedia.org/w/index.php?title=Evaluation_measures_(information_retrieval)&oldid=900146701, 2019, [Online; accessed 9-July-2019]. 5

[47] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105. 6

[48] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proceedings of the European Conference on Computer Vision*. Springer, 2014, pp. 818–833. 7

[49] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceeding of the IEEE International Conference on Computer Vision*. IEEE, 2017, pp. 2980–2988. 7

[50] Wikipedia contributors, "Rule of thirds — Wikipedia, the free encyclopedia," https://en.wikipedia.org/w/index.php?title=Rule_of_thirds&oldid=852178012, 2018, [Online; accessed 31-July-2018]. 7

[51] P. J. Huber *et al.*, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964. 7

[52] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256. 8

[53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. 8