Re-Compose the Image by Evaluating the Crop on More Than Just a Score

Yang Cheng
Purdue University
cheng159@purdue.edu

Qian Lin
HP Labs, HP Inc.
qian.lin@hp.com

Jan Allebach
Purdue University
allebach@purdue.edu

Abstract

Image re-composition has always been regarded as one of the most important steps during the post-processing of a photo. The quality of an image re-composition mainly depends on a person's taste in aesthetics, which is not an effortless task for those who have no abundant experience in photography. Besides, while re-composing one image does not require much of a person's time, it could be quite time-consuming when there are hundreds of images to be re-composed. To solve these problems, we propose a method that automates the process of re-composing an image to the desired aspect ratio. Although there already exist many image re-composition methods, they only provide a score to their predicted best crop but fail to explain why the score is high or low. Conversely, we succeed in designing an explainable method by introducing a novel 10-layer aesthetic score map, which represents how the position of the saliency in the original uncropped image, relative to that of the crop region, contributes to the overall score of the crop, so that the crop is not just represented by a single score. We conducted experiments to show that the proposed score map boosts the performance of our algorithm, which achieves a state-of-the-art performance on both public and our own datasets.

1. Introduction

Image re-composition is widely considered as one of the most important steps during the post-processing of a photo. Imagine a person who takes a photo using a mobile phone, which has an aspect ratio of 4:3, and wants to use it as the background on their personal laptop, which has an aspect ratio of 16:9. Imagine another person who is not good at taking high-quality photos and their photos consist of unwanted components in the peripheral area. In both cases, the person needs to use editing software, such as PhotoShop, to change the aspect ratio or improve the framing of the photo. An example of re-composing an image is shown in Figure 1, where the original image is cropped to have a better composition. The point of interest, the tree, now covers a larger area of the image, and is placed at a position that is closer to the center; all less important elements such as the bicycle and the smaller tree are removed, and the area occupied by the grass is significantly reduced. This process results in a much cleaner crop.

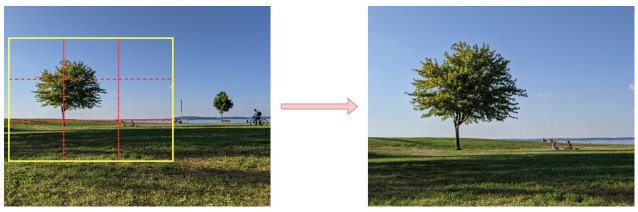

Figure 1. An example of re-composing an image. This process not only discards the bicycle and the smaller tree on the right that are considered less important, but also reduces the area occupied by the grass. The resultant crop is a much cleaner crop that focuses on the larger tree on the left.

With the emergence of smartphones, people are spending more time to take more photos but they are starting to find it difficult to post-process all of them; cropping an image is both time-consuming and tiring. Besides, re-composing an image is not an effortless task for those who have no abundant experience in photography, since the quality of the re-composition mainly depends on a person's taste in aesthetics. Therefore, for the purpose of both speeding up the process of re-composition and ensuring the high quality of the result, a method of automatically re-composing the images should be proposed.

In the past years, many methods have been introduced to re-compose the image; some earlier methods focus on locating the saliency region and deriving a crop based on this region [1, 13, 14], while the most recent and popular are the data-driven methods, which aim to learn the taste of professional photographers by training a model on the crops and scores evaluated by them [3, 15, 16, 17]. To facilitate the data-driven methods, several datasets have also been collected for research use [2, 3, 17, 19]. The datasets usually include the original image, and either its best crop or many crops of different quality with scores for each crop.

Although numerous methods have been proposed to re-compose images, some problems still exist with them. Tra-

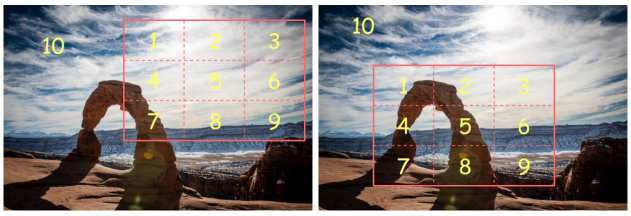

Figure 2. The crop in the right image is considered better than that in the left image, since the whole arch is included in the crop and its position in the crop is appropriate.

ditional methods do not require a large-scale dataset, but image re-composition is a task of high subjectivity, so the best crop derived from heuristics might not give enough aesthetic pleasure. Data-driven methods predict the best crop by learning from professional photographers' aesthetics. However, due to the limitation of public datasets where only crops and their scores are provided, they fail to explain why their predicted best crop is the most attractive one. Also, most of these methods did not consider how the positions of salient objects in the original image, either points of interest or distractions, relative to the position of the crop region, affect the aesthetics of the crop. For example, points of interest should be included in the crop while distractions should be excluded from the crop. Apart from the fact that the points of interest should be included in the crop, their positions in the crop are also very important; *e.g.*, in most cases, they should not be too close to the edge of the crop and sometimes simply placing them at the very center is also not the best decision.

To solve these problems, we put effort into enriching the ground truths so they are capable of showing whether a salient object is attractive or distracting, as well as its best location in the image. Also, we improve an existing network architecture by including an module to predict the enriched ground truths. The major contributions of this paper are

- A novel 10-layer aesthetic score map, which represents how the position of the saliency in the original uncropped image relative to that of the crop region contributes to the overall score of the crop.
- A neural network module that is inserted into an existing network architecture to predict the score map.

2. Related Work

2.1. Saliency-Driven Methods

Suh *et al*. [14] proposed to use the saliency map to preserve the important objects in thumbnails and to use face detection to ensure that the semantic information is not lost. Cheatle [1] noticed that the salient region might also include distractions, thus proposing a novel 3-class saliency map detection method that estimates the probability of the subjects, background, and distractions.

2.2. Data-Driven Methods

Chen *et al*. [3] proposed the View Finding Network (VFN) which is trained on pairs of crops to learn which one should be considered a better crop. Wei *et al*. [17] adopted the knowledge transfer framework to train their student model, the View Proposal Network (VPN), from the distilled knowledge extracted from their teacher model, the View Evaluation Network (VEN), which is trained on pairs of crops, similar to the VFN proposed by Chen *et al*. [3]. Therefore, the VPN simply needs to predict the scores of 895 pre-defined anchor boxes, thus making itself fast enough to run in real time. Wang and Shen [16] used a two-stage method: firstly, crop candidates are generated around the saliency predicted by their Attention Box Prediction (ABP) network, then the Aesthetics Assessment (AA) network is used to select the best crop. Tu *et al*. [15] proposed the idea of using the aesthetic score map, which is both composition-aware and saliency-aware, to predict the score of the crop. However, our aesthetic score map differs from theirs in that ours is generated directly from the ground-truth crops and scores.

3. Background

To better explain the ideas behind our method, we review two important concepts in this section: the rule of thirds and saliency. The rule of thirds is widely used by photographers when composing images, and the saliency represents the regions that attract people's immediate attention at first glance.

3.1. Rule of Thirds

The rule of thirds is a well-known rule that divides the image into thirds in both horizontal and vertical directions, and places points of interest at the intersections of lines or along the lines. This is illustrated in Figure 1. Inspired by this rule, we divide a crop into 9 regions to analyze the location of points of interest.

How the location of points of interest affects the aesthetics of the crop is explained in Figure 2. In each photo, the red box represents the crop, and the rule of thirds further divides the crop into 9 regions; the region that is outside the crop is denoted as the 10^{th} region. The crop in the image on the left in Figure 2 is considered bad, since part of the arch is not included in the crop and the part that is included and occupies Regions 4 and 7 should be probably shifted rightwards and upwards a little bit. However, the crop in the image on the right may be considered better than that in the image on the left; it covers the whole arch, the position of which in the crop is appropriate.

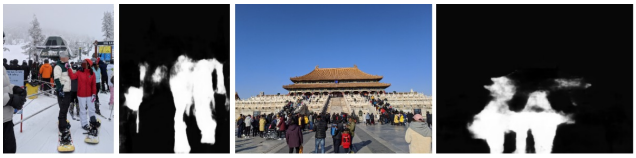

Figure 3. The first and third images from the left are the original images, and their predicted saliency maps are shown next to them. Notice that in the first original image, the hands of the first person from the left actually should be considered as a distraction, but are still considered salient by the saliency detection method. In the second original image, the crowd is considered salient although it is distracting.

3.2. Saliency

While a photographer is re-composing the image, they would always identify the points of interest first and then consider their appropriate positions in the crop. So it is not hard to imagine that an automatic re-composition method should take the same factor into consideration. Saliency detection has been one of the most popular tasks in computer vision for the past years, so many approaches have been introduced to accurately locate salient objects. The method that we adopt to locate the salient region was proposed by Hou *et al*. [7].

However, even though capable of fairly accurately locating the saliency region, this method lacks the ability to differentiate unwanted distractions from points of interest. Two examples are shown in Figure 3, where the distracting elements are considered salient. This is a rather reasonable and expected observation, since the distractions need to be salient enough to be distracting. This raises a problem with most saliency-based methods: the best crop derived from the saliency region is highly likely to include the distractions. In Section 4.2, we will explain how we solve this issue by introducing a novel 10-layer score map.

4. Methodology

To explain our method, in Section 4.1 we firstly introduce how our method predicts the best crop at a high level. Then, in Section 4.2, we introduce our novel 10-layer score map that can represent how the position of saliency in the original uncropped image relative to that of the crop region affects the score of the crop. In Section 4.3, we further illustrate how and why our crop evaluation model is designed in such a specific way that the positions of salient objects are taken into consideration. Lastly, training details and parameter settings are described in Section 4.4.

4.1. Image Re-Composition System

Our image re-composition system consists of two parts: crop candidates proposal and evaluation. In the training stage, the crop candidates are directly from the dataset, and their ground-truth scores are used to train the evaluation model. In the inference stage, the crop candidates are proposed using the sliding algorithm described below and evaluated by the trained evaluation model. This two-part system is not new, but has been used in the previous work [3, 17, 19]. Our major contributions lie in evaluating the crop candidates, where we design a novel score map that is capable of indicating whether a salient object is attractive or distracting and its best location in the crop, and propose a network module to predict this score map.

Crop Candidates Proposal

Given the desired aspect ratio, our method firstly generates crop candidates, represented by boxes, of different scales using the sliding window algorithm. Different scales, from 1.0 down to 0.5 with a step size of 0.05, of boxes are generated. A box of scale 1.0 is the maximum size rectangle with the desired aspect ratio that can fit in the image. For each scale, the box starts from the top left corner and slides rightwards and downwards. The number of crop candidates that can be generated depends on the desired aspect ratio and the aspect ratio of the original image.

Crop Candidate Evaluation

We propose a crop evaluation model that takes an original uncropped image and the coordinates of the crop candidates produced by the sliding window algorithm as input, and predicts their scores. The details of this model are discussed in Section 4.3. Finally, the crop candidate that receives the highest score is chosen as the best crop of the original image.

4.2. 10-Layer Aesthetic Score Map

Although the public dataset [17, 19] consists of thousands of different crops, their ratings are simply scores that range from 1 to 5, so they do not contain any information about why a crop receives a high or low score. This makes it harder for the evaluation model to learn the aesthetics of an image; similar features might be extracted by the model for two images of the same point of interest, but their ratings might be different due to a slight difference in the location of the point of interest. Also, existing data-driven methods are not explainable and are not capable of providing reasons why their predicted best crop is the most attractive one. Therefore, a method of enriching the ground truth should be proposed. When considering what factors decide the score of a crop, we noticed that it is the position and the size of salient objects that affect the overall score of the crop most.

Recall that we divide the original uncropped image into 10 different regions, where 9 of them are of equal size in the crop region and the 10^{th} region is the complement of

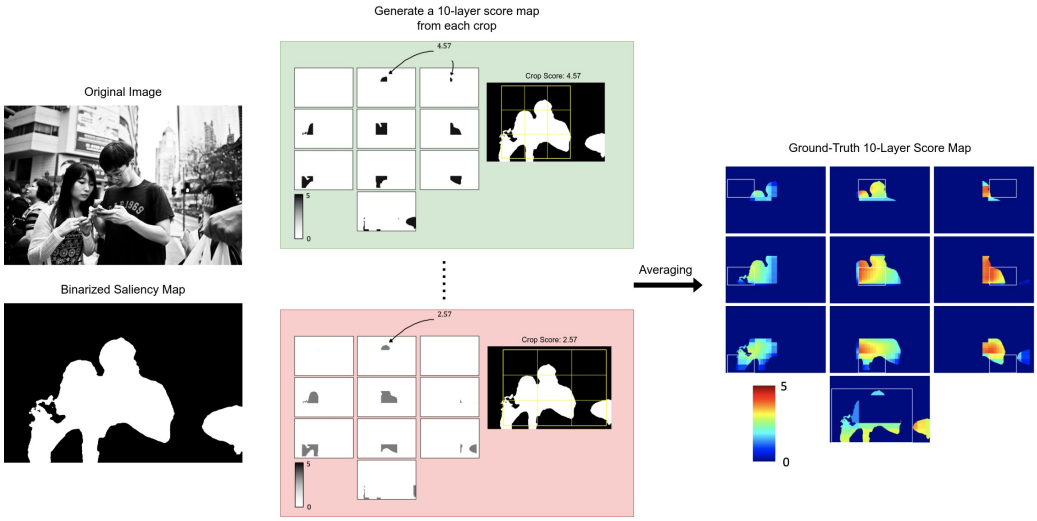

Figure 4. An illustration of the process of generating the ground-truth 10-layer score map for one image from the dataset [19]. To find the location of salient objects, a saliency detection method proposed by Hou *et al*. [7] is used and the saliency map is then binarized by Otsu's method [11] (first column). Two examples of generating the ground-truth 10-layer score map from a crop, its score and the saliency map are shown in the middle column. The final ground-truth 10-layer score map (third column) is obtained by averaging the ground-truth 10-layer score maps computed from all crops of an image in the dataset.

the crop region (Section 3.1). First of all, it is common sense that a distracting element should not be included in the crop, meaning that its salient region should be in the 10^{th} region, in that case, the score of the crop will be high. On the contrary, the salient region of a distracting element being in one of the 9 regions in the crop can lower the score of the crop. This thought can also be applied to an attractive element; its saliency region being completely or partially in the 10^{th} region can lower the score of the crop, while being inside the crop can increase the score, although it depends on which regions it covers.

Knowing whether a salient object is attractive or distracting and its position and size relative to the crop region can inform us whether the crop receives a high or low score. Conversely, we can be informed whether a salient object is attractive or distracting based on its position and size relative to the crop region and its score. For example, now suppose we have an uncropped image, a pair of a crop and its corresponding score. If the score is high and the majority of the saliency is outside the crop (or in the 10^{th} region), then probably this saliency is distracting. Similarly, if the score is low and the majority of the saliency is in the center of the crop (or in the 5^{th} region), then probably this saliency is distracting or should not be placed in the center.

4.2.1 Ground-truth 10-Layer Score Map Generation

With the idea of determining whether a salient object is attractive or distracting based on its location relative to the crop and the score of the crop, a score map can now be generated. The score map should consist of 10 layers because the salient object can belong to 10 different regions that are defined by different crops.

To find the location of salient objects, we use a saliency detection method proposed by Hou *et al*. [7] and binarize it using Otsu's method [11]. The location and score of the crop can be directly obtained from the dataset. We illustrate the process of generating the score map in Figure 4, where we use one image from the dataset [19] as an example. In the figure, the original image and binarized saliency map are both shown in the first column. In this image, the couple in the center is considered to be the point of interest, and the hand on the right is distracting, but both are considered salient by the saliency detection method. In the second column, two examples of how a 10-layer score map can be generated by a crop, its score, and the saliency map of the original image are demonstrated. To create the 10-layer score map, the original image is divided into 10 regions by the crop, and each pixel of the saliency map that belongs to the ith region is mapped onto the ith layer of the score map with a value of the score of the crop. In the example with the green background, the crop received a high score of 4.57 out of 5, most probably because the distracting hand is not included in the crop. It can be observed that the distracting hand receives a high score in the 10^{th} layer of the score map, and the couple receives a high score in the first 9 layers. In the example with the red background, a low score of 2.57 is assigned to the crop due to the inclusion of the distracting hand. As it can be seen, the majority of the hand is in the 9^{th} layer, and the values are smaller.

Since the dataset contains nearly 100 crops for each image, the final 10-layer score map can be computed by av-

eraging all the score maps generated by each crop. This is shown in the third column of Figure 4. From the final score map, it can be observed that the distracting hand receives high values in the 10th layer, while receiving low values in the 9th layer, and is not activated in the rest of the layers. This indeed validates the purpose of the score map, which is to show the location where a salient object should be placed to maximize the score of the crop. In this example, the hand receives high values in the 10th layer, so it is considered distracting and should be excluded from the crop to maximize its score.

The procedure is described in Algorithm 1. Here, we define the variables in detail.

- The original uncropped image is of size $H \times W$. The binarized saliency map, denoted as $\boldsymbol{S}$, is of size $H \times W$ and has values either 0 or 1 at each pixel.
- $crop\_region$ defines the crop region in the original image. More specifically, if we let (y_1, x_1) be the coordinate of top left corner and (y_2, x_2) be the coordinate of bottom right corner, then $\boldsymbol{A}_{crop\_region}$ is defined as $(\boldsymbol{A}_{ij})_{\substack{y_1 \le i < y_2 \\ x_1 \le j < x_2}}$, where $\boldsymbol{A}$ is an arbitrary matrix. The ten regions introduced in Section 3.1 can then be represented as nine subregions and the complement of A_{crop_region} according to the following:

$$\boldsymbol{A}_{crop\_region_1} \triangleq (\boldsymbol{A}_{ij})_{\substack{y_1 \le i < y_1 + \frac{1}{3}(y_2 - y_1) \\ x_1 \le j < x_1 + \frac{1}{3}(x_2 - x_1)}} \tag{1}$$

$$\boldsymbol{A}_{crop\_region_2} \triangleq (\boldsymbol{A}_{ij})_{\substack{y_1 + \frac{1}{3}(y_2 - y_1) \le i < y_1 + \frac{2}{3}(y_2 - y_1) \\ x_1 + \frac{1}{3}(x_2 - x_1) \le j < x_1 + \frac{2}{3}(x_2 - x_1)}} \tag{2}$$

$$\dots$$

$$\boldsymbol{A}_{crop\_region_{10}} \triangleq \overline{(\boldsymbol{A}_{ij})_{\substack{y_1 \le i < y_2 \\ x_1 \le j < x_2}}} \tag{3}$$

- The output 10-layer aesthetic score map is represented by a matrix $\boldsymbol{M}^{10 \times H \times W}$, and $\boldsymbol{M}[k]_{crop\_region_k}$ represents the subregion of $\boldsymbol{M}[k]$ that is defined by $crop\_region_k$, which is the kth region of the $crop\_region$, $k \in \{1, 2, \dots, 10\}$.

4.2.2 Score Map Interpretation

In Section 4.2.1, the method of generating a ground-truth 10-layer score map for the original image was introduced, but for an unknown given crop, how do we check if the salient objects are at their best location? It turns out that we can simply divide the original image into 10 regions based on this crop, select Region i in the ith layer of the score map, and then concatenate these 10 regions together to form a crop-specific heatmap, which can be used to visualize the quality of the crop. This is illustrated in Figure 5. In the example with the blue background, the crop is considered a good crop as it does not contain any distracting element, and it can be observed that the saliency of both the couple and the hand has large values in the crop-specific heatmap. This indicates that it is better to include the couple in the crop while excluding the hand. However, in the example with the purple background, the values of the crop-specific heatmap are much lower in the saliency of both the couple and the hand. This means that they are not at their best location. The position of the couple is too close to the left edge of the crop and the distracting hand is included in the crop.

Algorithm 1: 10-Layer Score Map Generation

Input: Binarized saliency $\boldsymbol{S}$, a list of crop regions and their scores $(crop\_region, score)$

$\boldsymbol{M}$ = **zeros**(10, H, W), $\boldsymbol{C}$ = **zeros**(10, H, W)

foreach $(crop\_region, score)$ **do**

  for *k=1,2,...,10* **do**

    $\boldsymbol{M}[k]_{crop\_region_k} := \boldsymbol{M}[k]_{crop\_region_k} + score * \boldsymbol{S}_{crop\_region_k}$

    $\boldsymbol{C}[k]_{crop\_region_k} := \boldsymbol{C}[k]_{crop\_region_k} + \boldsymbol{1}$

  end

end

`/* Average the score map */`

$\boldsymbol{M} := \boldsymbol{M}/\boldsymbol{C}$;

Output: 10-layer score map $\boldsymbol{M}$

4.3. Crop Evaluation Model

The architecture of our proposed model is illustrated in Figure 6. The inputs are simply the original uncropped image and the coordinates of the crop, represented by the red box. And the outputs are the 10-layer score map of the original uncropped image and the score of the crop on a scale of 1 to 5. The former one is an intermediate output that can guide the training of the network and provide a visual explanation; the latter one is the final output that is used to find the best crop.

Following the architecture used by Zeng *et al*. [19], we adopt MobileNet-V2 [12] as the backbone of our network to extract the feature maps. Feature maps at two different scales are concatenated together to deal with different sizes of points of interest. Besides the feature map extraction, one of the intermediate feature maps is fed into an hourglass network to predict the 10-layer aesthetic score map. The hourglass network, proposed by Newell *et al*. [10], is an encoder-decoder network that has been commonly used to predict the heat maps of joints in the human body for the task of human pose estimation.

Subsequently, both RoI Align and RoD Align operators are applied to both the concatenated feature maps and the

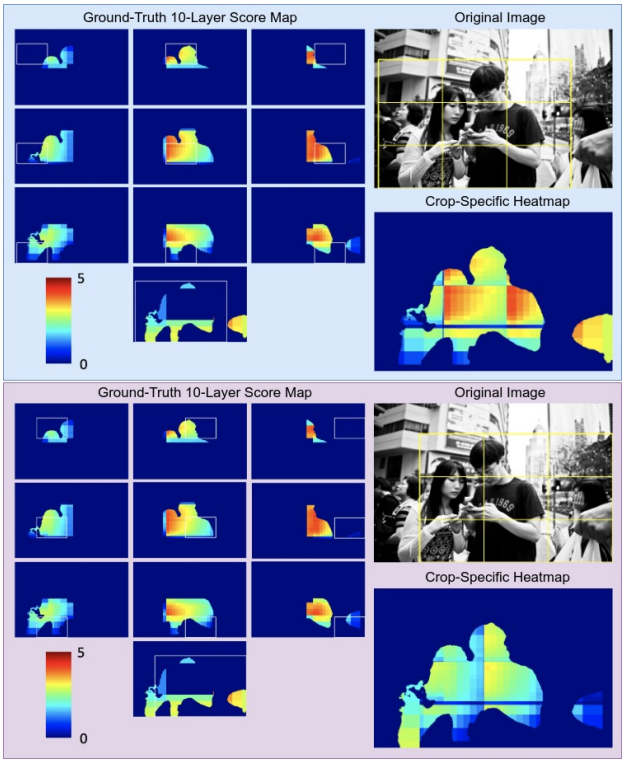

Figure 5. Two examples of how to use the ground-truth 10-layer score map to visualize the quality of a random crop. For each example, the crop divides the image into 10 regions, then the final crop-specific heatmap (bottom right) can be formed by concatenating Region i from ith layer (white boxes) for $i = 1, 2, 3, \ldots, 10$. As it can be observed, the crop that does not include the distracting hand results in a heatmap where the values in both regions (the couple and the hand) are high. This indicates that both salient objects are at their best location.

predicted 10-layer score map. The resulting outputs are then concatenated together, and a score is predicted by the remaining part of our network.

4.3.1 RoI Align and RoD Align

RoI Align [6] was initially proposed to remove the quantization error that exists in RoI Pooling [5] with object detection tasks. However, they share the same objective: mapping the region in the original image onto the feature maps and extracting these features. Zeng *et al*. [19] utilized this idea in their network to map the crop region in the original uncropped image onto the feature maps. In addition, considering that the region not included in the crop cannot be ignored either, they also proposed RoD Align, based on the idea of RoI Align. Opposite to the RoI, or the Region of Interest, it is the region that should be discarded, and is named the RoD, or the Region of Discard. It is illustrated in the yellow region in Figure 6. The RoI Align operator is applied on the feature map to extract the features within the crop region. The RoD Align operator includes zeroing out the features within the crop region and mapping the feature map to another one of the same size as that of the output of the RoI Align operator for later concatenation.

As introduced in Section 4.2.2, given a crop and the 10-layer score map, a crop-specific heatmap can be generated to show whether the salient objects are at their best location relative to the crop region. In the evaluation model, we follow the same procedure, but use the RoI Align operator to extract the features within the first 9 regions from the 10-layer score map and group them together as a 3×3 grid. Then we use the RoD Align operator to zero out the features within the crop region from the score map and map it to another one so it can be concatenated with the output from the RoI Align operator. This operation is illustrated in the green region in Figure 6.

4.4. Parameter Settings

To train our network to predict the 10-layer aesthetic score map and the score of the crop accurately, we use the Mean Squared Error (MSE) loss and the Smooth L1 loss, or the Huber loss. The learning rate is set to 10^{-4} and the Adam optimizer [8] is used. The shorter side of the image, either width or height, is resized to 256 before being fed into the network.

5. Experiments

5.1. Performance on a Public Dataset

GAIC Dataset [19] is a public dataset that consists of 1237 images. Each image contains approximately 90 crops rated by 19 recruited subjects who had passed a test on photography composition. The score of each crop is on a scale of 1 to 5, aka a Mean Opinion Score (MOS). For each image, we use the saliency detection proposed in [7] to generate the saliency map, which is then binarized by Otsu's method [11], and lastly we generate the 10-layer score map according to Algorithm 1. Out of these 1237 images, 1037 images were used to train our model, while the remaining images are for testing our model.

Following [19], we use the three metrics defined below to compare the performance of existing methods on the GAIC testing dataset.

- **Top N Accuracy** measures the proportion of the time when the true best crop is among the top N predicted best crops.
- **Pearson Correlation Coefficient (PCC)** measures the correlation between the predicted and true scores of a list of crops.
- **Spearman's Rank Correlation Coefficient (SRCC)** measures the correlation between the ranks of the predicted and true scores of a list of crops.

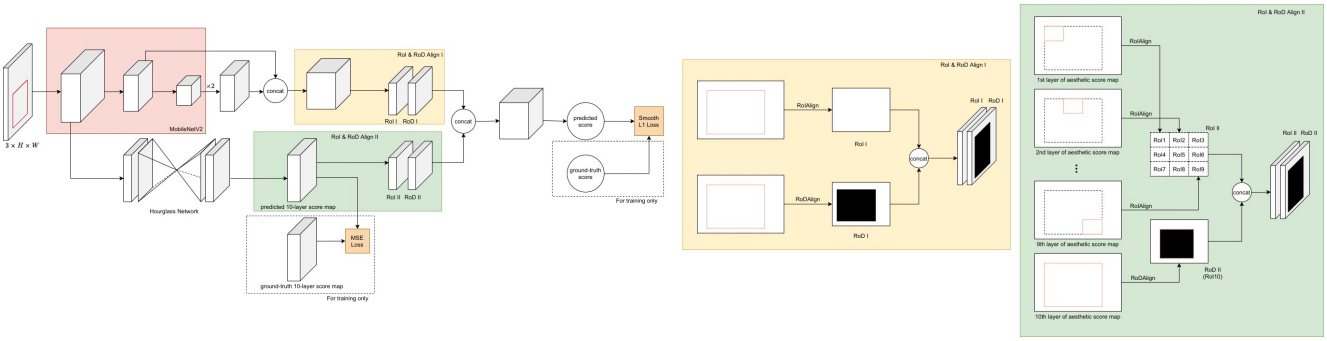

Figure 6. An illustration of our network. MobileNetV2 is used as the backbone to extract the feature maps (red region). After necessary upsampling, represented by $\times 2$, two feature maps at different scales are concatenated together. Also, an hourglass network predicts the 10-layer aesthetic score map. Then RoI Align and RoD Align are applied on both the feature map (yellow region) and the score map (green region). More detailed illustration of RoI Align and RoD Align operators can be seen on the right. Finally, the score of the crop is predicted by the network.

	$Acc_{1/5}\uparrow$	$Acc_{1/10}\uparrow$	$PCC\uparrow$	$SRCC\uparrow$
Li *et al*. (A2-RL) [9]	23.0	38.5	-	-
Wei *et al*. (VPN) [17]	40.0	49.5	-	-
Chen *et al*. (VFN) [3]	27.0	39.0	0.470	0.450
Wei *et al*. (VEN) [17]	40.5	54.0	0.653	0.621
Tu *et al*. (ASM-Net) [15]	54.3	71.5	-	0.766
Zeng *et al*. (GAIC) [19]	62.5	78.5	**0.806**	0.783
Xu *et al*. [18]	21.0	32.0	0.459	0.432
Ours	**66.0**	**83.0**	**0.806**	**0.787**

Table 1. Performances of different methods on GAIC testing dataset.

The performances of 8 different cropping systems on the testing images are reported in Table 1. Among these methods, ours achieves the best performance. It is worth noting that the biggest difference between GAIC and ours is the branch that we inserted to predict the 10-layer score map in the network structure we developed. A significant improvement can be observed, thus indicating the value of the 10-layer score map.

Figure 7. The first images from the left are the original images, whose width is larger than the height, followed by the crops predicted by 5 different cropping algorithms. The value below each crop represents the percentage of subjects who liked the crop.

5.2. Performance on Our Own Dataset

Image cropping is a task of high subjectivity, so the best and most straightforward way of evaluating the performance of a cropping method is to ask subjects whether

Figure 8. The first images from the left are the original images, whose width is smaller than the height, followed by the crops predicted by 5 different cropping algorithms. The value below each crop represents the percentage of subjects who liked the crop.

or not they like the crops generated by it. Therefore, to do so, we collected 100 images, where 50 images contain at least one obvious main object (e.g., person, animal) and the remaining images are pure landscape photos. 17 subjects participated in the experiment. For each image, they were asked to either like or dislike the crop generated by each of 5 different cropping algorithms, which are shown in the first column of Table 2. To prevent the subjects from being tired of the experiment and producing low-quality ratings, we kept the duration of the experiment within one hour by selecting only 4 cropping algorithms including ours from the algorithms listed in Table 1. Another method proposed by Cheatle [1] is also included in the subjective experiment. However, it is not evaluated on the GAIC testing dataset due to its design of the cropping algorithm: it predicts the best crop of a fixed, pre-defined aspect ratio by expanding a box from the minimum crop area, thus making it unsuitable for being evaluated on the metrics $Acc_{1/N}$, *PCC*, and *SRCC*. Therefore, it is not included in Table 1.

To remove any possible outlier that could affect our results, for each cropping method, we applied the 1.5 IQR rule [4] on the number of likes received by the method to

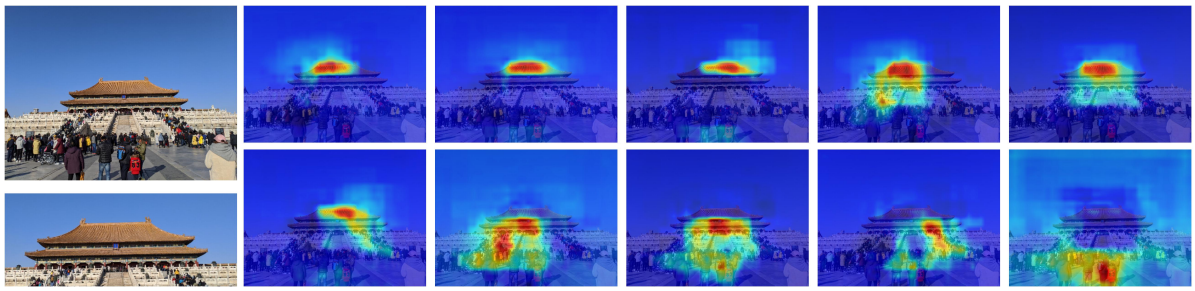

Figure 9. **Top left:** the original image, **Bottom left:** the predicted best crop, **Remaining images:** the 10 layers of the score map predicted by our model. Notice that salient region of the palace is only activated in the first 9 layers, and the salient region of the distracting crowd is only activated in the 10^{th} layer, indicating that our model is capable of distinguishing between attractive and distracting elements.

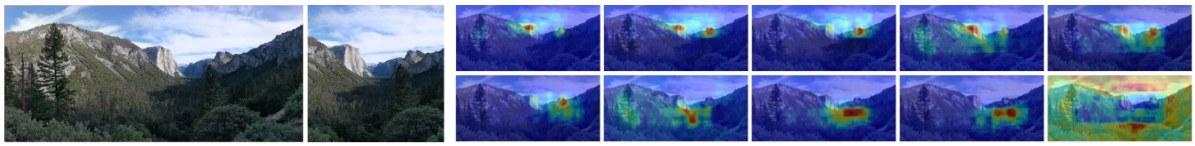

Figure 10. **First image from the left:** the original image, **Second image from the left:** the predicted best crop, **Remaining images:** the 10 layers of the score map predicted by our model. Our method is capable of finding the main point of interest, which is the rock face, from this panoramic image, and predicting a reasonable best crop.

identify which subjects are outliers. If a subject is considered as an outlier for at least one method, we remove all the data points of this subject from our results.

After conducting the data analysis, we removed 3 subjects whose data points were detected as outliers. Table 2 summarizes the results by showing the average and standard deviation of number of likes received by each method. Our method still outperforms all other methods by receiving the highest number of likes, while the standard deviation is still low. Also, notice that the performance results coincide with those in Table 1: the predictions by VFN are not reliable, while our method outperforms all other methods.

	Mean ↑	**S.T.D.** ↓
Cheatle [1]	77.29	**4.71**
Chen *et al*. (VFN) [3]	26.64	5.84
Zeng *et al*. (GAIC) [19]	79.57	7.09
Xu *et al*. [18]	59.14	9.15
Ours	**87.57**	4.88

Table 2. Average and standard deviation of number of likes received by 5 different algorithms.

Figures 7 and 8 show some examples of crops predicted by these 5 cropping algorithms. The value below each crop represents the percentage of subjects who liked the crop. With the help of the 10-layer aesthetic score map, our method is capable of placing the important saliency (points of interest) in the correct positions, and excluding the distracting saliency (distractions).

5.3. Visualization of 10-Layer Score Map

To show that our cropping algorithm is explainable, we provide predicted 10-layer score maps for two images in Figures 9 and 10. For example, in Figure 9, the salient region of the palace is activated in the first 9 layers of the score map but not in the 10^{th} layer, showing that our model is capable of locating the point of interest. The salient region of the crowd is only activated in the 10^{th} layer, showing that our model correctly predicts the crowd as a distraction. This indicates that our model is capable of distinguishing between attractive and distracting elements. In Figure 10, our method is also able to find the main point of interest, which is the rock face, from the panoramic image, and predict a reasonable best crop.

6. Conclusion

In this paper, we introduced a novel 10-layer score map, which represents how the position of the saliency in the original uncropped image relative to that of the crop region contributes to the overall score of the crop. One disadvantage of direct usage of public datasets is that each crop is only evaluated by a single score, so it is hard to explain why this crop receives a high or low score, thus making it hard to train the network to learn the quality of the crop. Conversely, our proposed 10-layer score map allows evaluating a crop on more than just a score. The experiments show that our method outperforms all the other methods in terms of different metrics. Also, our method is easily explainable: the predicted 10-layer score map is capable of showing why our model assigns a high or low score to a crop.

References

[1] Phil Cheatle. Automatic image cropping for republishing. In *Proceedings of SPIE*, volume 7540, pages 75400O–75400O–9, Springfield, VA, 2010. IS&T-The Society for Imaging Science and Technology.

[2] Y. Chen, T. Huang, K. Chang, Y. Tsai, H. Chen, and B. Chen. Quantitative Analysis of Automatic Image Cropping Algorithms: A Dataset and Comparative Study. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 226–234, 2017.

[3] Yi-Ling Chen, Jan Klopp, Min Sun, Shao-Yi Chien, and Kwan-Liu Ma. Learning to Compose with Professional Photographs on the Web. In *ACM Multimedia 2017*, 2017.

[4] F.M Dekking. *A Modern Introduction to Probability and Statistics Understanding Why and How*. Springer Texts in Statistics. 1st ed. 2005. edition, 2005.

[5] R. Girshick. Fast R-CNN. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

[6] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[7] Q. Hou, M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr. Deeply Supervised Salient Object Detection with Short Connections. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5300–5309, 2017.

[8] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[9] Debang Li, Huikai Wu, Junge Zhang, and Kaiqi Huang. A2-RL: Aesthetics Aware Reinforcement Learning for Image Cropping. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8193–8201, 2018.

[10] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked Hourglass Networks for Human Pose Estimation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 483–499, Cham, 2016. Springer International Publishing.

[11] N. Otsu. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

[12] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520, 2018.

[13] F. Stentiford. Attention Based Auto Image Cropping. In *ICVS 2007*, 2007.

[14] Bongwon Suh, Haibin Ling, Benjamin B. Bederson, and David W. Jacobs. Automatic thumbnail cropping and its effectiveness. In *Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology*, UIST '03, page 95–104, New York, NY, USA, 2003. Association for Computing Machinery.

[15] Yi Tu, Li Niu, Weijie Zhao, Dawei Cheng, and Liqing Zhang. Image cropping with composition and saliency aware aesthetic score map. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12104–12111, Apr. 2020.

[16] W. Wang and J. Shen. Deep Cropping via Attention Box Prediction and Aesthetics Assessment. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2205–2213, 2017.

[17] 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 *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5437–5446, 2018.

[18] Shaoyuan Xu. *Content Understanding for Imaging Systems: Page Classification, Fading Detection, Emotion Recognition, and Saliency-Based Image Quality Assessment and Cropping*. Ph.D. Dissertation, Purdue University, West Lafayette, IN 47906, 2020.

[19] Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Grid Anchor based Image Cropping: A New Benchmark and An Efficient Model. *CoRR*, abs/1909.08989, 2019.