Computational Visual Media
https://doi.org/10.1007/s41095-021-0263-3 Vol. 9, No. 1, March 2023, 87–107
Research Article

Focusing on your subject: Deep subject-aware image composition recommendation networks

Guo-Ye Yang[1], Wen-Yang Zhou[1], Yun Cai[1], Song-Hai Zhang[1] (✉), and Fang-Lue Zhang[2]

Abstract Photo composition is one of the most important factors in the aesthetics of photographs. As a popular application, composition recommendation for a photo focusing on a specific subject has been ignored by recent deep-learning-based composition recommendation approaches. In this paper, we propose a subject-aware image composition recommendation method, SAC-Net, which takes an RGB image and a binary subject window mask as input, and returns good compositions as crops containing the subject. Our model first determines candidate scores for all possible coarse cropping windows. The crops with high candidate scores are selected and further refined by regressing their corner points to generate the output recommended cropping windows. The final scores of the refined crops are predicted by a final score regression module. Unlike existing methods that need to preset several cropping windows, our network is able to automatically regress cropping windows with arbitrary aspect ratios and sizes. We propose novel stability losses for maximizing smoothness when changing cropping windows along with view changes. Experimental results show that our method outperforms state-of-the-art methods not only on the subject-aware image composition recommendation task, but also for general purpose composition recommendation. We also have designed a multi-stage labeling scheme so that a large amount of ranked pairs can be produced economically. We use this scheme to propose the first subject-aware composition dataset SACD, which contains 2777 images, and more than 5 million composition ranked pairs. The SACD dataset is publicly available at https://cg.cs.tsinghua.edu.cn/SACD/.

Keywords subject-aware image composition; image cropping; deep learning; recommendation

1 Introduction

The rapid recent advances in digital technology make it possible for the public to capture and produce photos with comparable resolution and sharpness to those taken by professional equipment. However, most photos taken by novice users have poor aesthetic quality with regard to composition due to their lack of photographic skills. Researchers have investigated computational approaches to suggest to ordinary users the best crop of the original photo with highest aesthetic quality. Although composition evaluation of a given crop is a subjective process, prior research has acknowledged that it follows some objective laws which can be learned by deep neural networks [1, 2]. Based on the learned implicit features for composition evaluation, some deep architectures have achieved great success at optimal cropping window prediction [3, 4].

In photo composition theory and practice, it is widely accepted that good photos generally satisfy principles related to their main subject, such as directing attention to the subject, and removing objects that distract attention from the subject [5]. Moreover, ordinary users usually focus on specific subjects when they take photos. A good composition should put the main subjects in a place where they are a dominant part of the image. However, existing

1 BNRist, Department of Computer Science and Technology, Tsinghua University, Beiing 100084, China. E-mail: G.-Y. Yang, yanggy19@mails.tsinghua.edu.cn; W.-Y. Zhou, zhouwy19@mails.tsinghua.edu.cn; Y. Cai, cynthiacai1107@gmail.com; S.-H Zhang, shz@tsinghua.edu.cn (✉).

2 School of Engineering and Computer Science, Victoria University of Wellington, Wellington 6012, New Zealand. E-mail: fanglue.zhang@vuw.ac.nz.

Manuscript received: 2021-08-17; accepted: 2021-11-01

清华大学出版社 Tsinghua University Press Springer

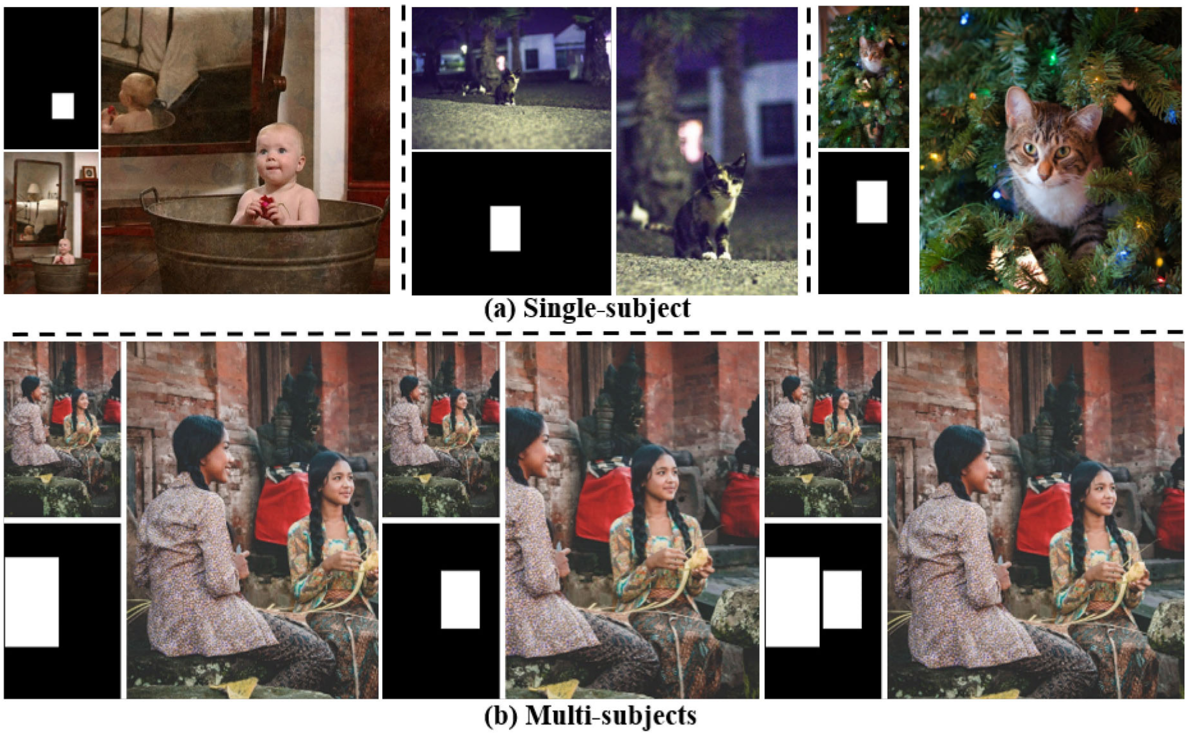


Fig. 1 Examples of SAC-Net. (a) Cropping examples for single-subject images. Given an original image and a subject's window mask, our method can recommend a sub-image with high aesthetic value. (b) Cropping examples for multi-subject images. Different composition results are obtained for different input subject window masks.

deep-learning-based composition recommendation methods have not explicitly taken the main subject of an image into consideration, and consequently achieve unsatisfactory results, especially in scenes containing multiple objects. To fill that gap in the literature, we propose a subject-aware composition recommendation method based on deep neural networks. Our model takes an RGB image and a subject window mask as input, and returns cropping recommendations for good compositions containing the subject. To support the training of subject-aware models, we also propose a dataset containing manually annotated cropping windows whose composition qualities consider their main subjects.

Many of the composition rules commonly used in photography rely on the spatial layout of objects. When evaluating the visual quality of photos, users are very sensitive to even minor changes in spatial information such as object positions in the photo and spaces between objects and photo boundaries. Thus, the accuracy of the predicted cropping windows is of critical importance. However, most previous deep-learning-based methods predict optimal cropping windows by evaluating several pre-set windows on the original photo [2–4, 6], which inevitably misses the best cropping in many cases. To address this issue, we are the first to propose a method that can directly regress cropping windows with arbitrary positions and sizes to give the optimal composition. We propose a deep *subject-aware composition recommendation network* (SAC-Net), which has two stages: (i) predict candidate scores for all possible coarse cropping windows, and (ii) recommend high score windows by refining their corner points and regressing their final scores.

In SAC-Net, the basic structure enabling the generation of cropping windows with arbitrary positions and sizes is inspired by the object detection network, Faster R-CNN [7]. However, it is non-trivial to directly apply an object detection network to the composition task for the following fundamental reason: unlike the object detection task in which the output box just has an object label, the image cropping task requires the model to have the ability to provide an aesthetic score for predicted boxes. Therefore, we introduce the following learning and prediction paradigms for the task of subject-aware automatic cropping. Firstly, to find the optimal crop among all possible windows, we adopt a coarse candidate score map and carefully design loss functions to allow the network to learn to evaluate the aesthetic quality of different boxes, including a continuity loss to maximize smoothness when cropping windows undergo minor changes and a score distance loss for learning the candidate score map. Secondly, since the final aesthetic score depends on the box size and position refined by regression, we propose a different inference pipeline in which we sequentially connect the sub-modules of the final box regression and score evaluation during the inference stage. Thirdly, we also improve the anchor selection and feature map organization schemes to adapt them to the cropping recommendation task. Our experimental results show that the above paradigms work well on the proposed interactive composition task: our method significantly outperforms the state-of-the-art methods and the alternatives of object detection networks.

To support the training of our deep model, we need a dataset containing manually annotated cropping windows with their subject-aware composition qualities. Since the existing datasets [2, 3] are labeled without considering the main subjects of the original photos, we have built the first large-scale subject-aware image composition dataset (SACD). It is non-trivial to properly label composition quality for a large number of cropping windows at a reasonable cost. Wei et al. [2] and Zeng et al. [3] selected some windows, and asked artists to judge their composition quality pair by pair. But their strategy could miss some good cropping windows, which may confusion a network. Also, it would be too expensive for artists to label all possible cropping window pairs. Thus, we design a multi-stage filtering annotation scheme enabling annotators to effectively evaluate the composition quality of all possible cropping windows with the specified subject. A large number of ranked pairs can be produced through our labeling process. Finally, we label 2777 images and achieve more than 5 million effective composition ranked pairs. Each image contains at least 8 crops with good composition for one subject.

Our main technical contributions are thus:

1. A subject-aware image composition recommendation method, SAC-Net. Based on its two-stage network architecture with carefully designed learning paradigms, it is able to directly predict accurate cropping windows without using any pre-set windows. It substantially outperforms the state-of-the-art methods in the proposed interactive setting.
2. New learning schemes and losses to enable such a two-stage network to provide proper cropping windows and evaluate their aesthetic quality.
3. The first subject-aware composition dataset, SACD, which contains 2777 images, more than 24,000 cropping windows, and more than 5 million composition ranked pairs.

2 Related work

2.1 Photo composition methods

Researchers have demonstrated that photo composition can be computationally evaluated and improved. Earlier work mostly uses manually designed feature extraction [8–14], composition scoring functions [15–20] such as the amount of salient content [21, 22], semi-automatic composition based on eye tracking [23], training a composition evaluation model such as a support vector machine (SVM) [24–26], or matching to a well-composed template library to optimize composition [27]. These methods have achieved good results on simple pictures, but appear less adaptable to more complex pictures.

In recent years, many deep-learning-based methods [28–31] have been proposed to learn the aesthetic value of pictures by training convolution neural networks (CNN), and have achieved good results on tasks related to photo composition. Ref. [28] proposes a real-time composition recommendation algorithm based on feature fitting using the user's favorite image set. Ref. [29] uses hand-crafted features and SVMs to learn the composition quality of cropping windows, and then exhaustively enumerates windows and selects the one with the highest score during use. A dataset containing 1000 images was constructed in their work. Ref. [30] modifies the CNN network structure and proposes layers such as local contrast normalization, which can effectively extract image quality features. Ref. [31] inputs global and local information for classification. Later, more sophisticated networks [32–35] were designed to extract richer feature information from pictures.

Ref. [32] proposes a deep multi-patch aggregation (DMA) network, using only image patch blocks as input. Ref. [33] explores more aesthetic attributes, including the three-point line, whether the content is interesting, etc. Ref. [34] proposes an adaptive spatial pooling layer, which can handle input images of different sizes and aspect ratios, thus avoiding the problem of image distortion. Ref. [35] directly predicts the regression box based on the target detection framework. At the same time, it provides a dataset with 28,064 images and 70,048 annotations. The above methods are mostly proposed for assessing aesthetic qualities of given images, and are not directly applicable to recommend good cropping windows.

More recently, researchers have paid more attention to cropping window prediction for good composition. Refs. [36] and [37] first propose the use of deep learning in the composition task with a two-stage composition method consisting of attention box prediction and aesthetic assessment, with achieved good results. Ref. [38] performs saliency estimation to generate thumbnails for stereoscopic photo pairs. Ref. [1] assumes that the original image taken by the photographer should score higher in composition than the cropped image, when designing the ranking loss for network training. Ref. [39] finds the best cropping window by reinforcement learning. Ref. [40] proposes a method based on comparison of the qualities of two compositions, which is more accurate than direct scoring. Wei et al. [2] first proposed use of a comparison-based dataset to train the neural networks. They proposed a view evaluation network (VEN) and a view proposal network (VPN) network, which greatly improves the performance. Based on recent advances in deep neural networks, many researchers in this field have worked on improving network structure to provide better results [3, 4, 6, 41–43]. Other researchers have also provided dataset labeling schemes [3] and novel losses [44] to help the learning process in composition recommendation networks. In particular, Ref. [6] proposes the composition- and saliency-aware ASM-Net, which can learn the internal mechanism of composition to a certain extent. Ref. [41] proposes a network that can aesthetically score full-resolution images. Ref. [43] proposes an end-to-end, composition recommendation algorithm based on saliency maps. Ref. [45] proposes to use distribution dissimilarity between high quality images and cropped images to train the composition model, instead of using training images with ground truth cropping. This work also proposed a saliency loss to make the model focus more on the salient parts of the image, but does not let the user choose the target subject. Ref. [46] collectes a dataset with 51k images, and 5 crops with different aspect ratios are annotated for each image; its CNN model can directly obtain composition results for 5 aspect ratios. Ref. [47] uses a key composition map (KCM) to encode the composition rules and built a network that can explicitly apply the learned rules.

The previous methods have achieved good results for the composition task, but the importance of subjects is neglected in the above methods, which sometimes leads to poorly located subjects. Furthermore, even after choosing the best composition containing the subject generated by these methods, there may still be a chance that other subjects are the visual center of the image, while the selected subject is ignored; there is no use of secondary objects to highlight the subject in a harmonious composition.

In addition, the effectiveness and efficiency of the previous methods [2–4] depend on the number of preset cropping windows to some extent. Our submodule for generating candidate score maps can provide a smaller number of valuable crops for later accurate cropping refinement and score regression. It significantly improves results and can provide state-of-the-art performance on public datasets.

2.2 Photo composition datasets

Several datasets exist for evaluating composition quality. Some [48–52] focus on the aesthetic score of a single picture. These datasets enable researchers to solve the composition recommendation problem through deep learning. Ref. [48] collects online peer-rated photos as a dataset, giving the average score, number of downloads, and number of ratings. Ref. [49] provides a dataset containing 17,613 images with manually annotated aesthetic scores. Ref. [50] presents the AVA database, which contains more than 250,000 pictures with aesthetic scores. Ref. [51] presents the FLMS dataset, where each picture contains 10 hand-labeled excellent composition frames. Ref. [52] proposes a dataset for selecting the best photos amongst similar photos, and annotates the partial order of aesthetic quality for each pair of similar photos.

In 2018, Wei et al. [2] proposed a comparison-based composition (CPC) dataset, which contains 10,800 images, with 24 views for each image and more than one million ranked pairs for different crops of the same image. They show that comparison-based composition datasets are more conducive to the composition task. In 2019, Zeng et al. [3] presented a grid-anchor-based image cropping dataset (GAICD). They provide up to 90 fixed crops for each image and score each crop. These datasets are both annotated in terms of the entire picture, which does not fit the subject-aware cropping recommendation task. Even if cropping windows that do not contain the subject are discarded, the remainder are not evaluated with the consideration of the subject. They thus can not be directly used in our task.

3 The SACD dataset

3.1 Overview

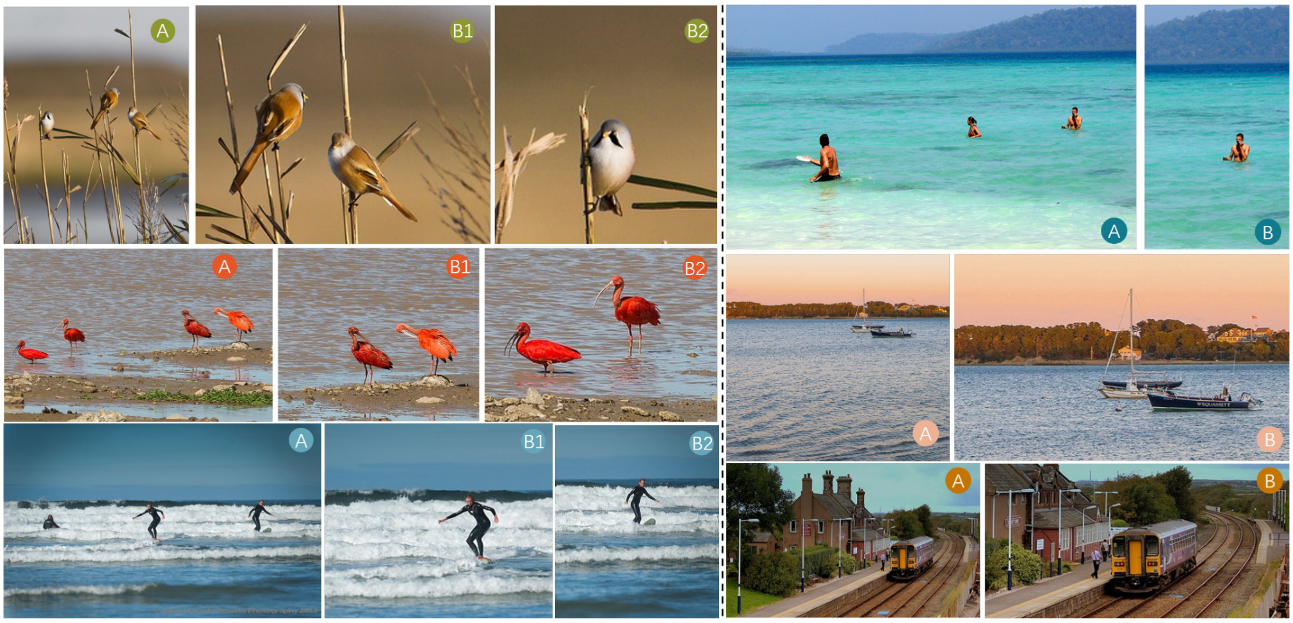
We present here a subject-aware composition dataset (SACD) for the task of image cropping for composition. It includes 2777 images and 5.2 million ranked pairs. The labeled cropping windows all have high aesthetic value with a certain focused subject. Some example images of our dataset are shown in Fig. 2. Rather than undirectedly finding cropping windows with good composition for the entire image, models trained on SACD can give cropping recommendations for a specific subject in the image.

3.2 Image collection

All of our images were collected from the Microsoft Common Objects in Context (MS COCO) dataset [53], which consists of images annotated with object detection information. All MS COCO images were collected from Flicker, a website for amateur photographers to upload their photographs. We use MS COCO images to build our dataset for two reasons. Firstly, most Flicker images were taken in real photographic environments, enabling the trained model to work better in real applications. Secondly, most were not taken by professionals, and leave large room for composition improvement. We retain images based on two rules: firstly, at least one subject should be included in the image, and secondly, the subject size should not be too large or too small.

3.3 Image annotation pipeline

After collecting eligible images from MS COCO, we invited professional artists to label each image in a comprehensive process. To ensure the diversity and quality of the labeled results, our labeling process has four stages.



(a) Multi-subjects

(b) Single-subject

Fig. 2 Subject-aware composition dataset (SACD) examples. (a) Images with multiple subjects. Original images are tagged A, and sub-images labeled by professional artists for different subjects are tagged B1 and B2. (b) Images with a single subject, with tags A and B having the same meaning as in (a).

1. We assign a subject for each picture, such as a person, a building, a bird, etc.; all annotations should be focused on this subject. Subject detection is done by Mask R-CNN [54]. We divide the images into 5 × 5 image blocks on average, and then select ij $(2 \leqslant i \leqslant 5,\ \ 2 \leqslant j \leqslant 5)$ consecutive image blocks from top to bottom and left to right as candidate images. In this way, 100 sub-images are generated with different aspect ratios and sizes, covering different regions of the original image. At this stage, the annotator looks through the 100 sub-images, and selects at least 20 sub-images for the next stage.
2. The annotator now carefully goes through the images from the previous stage, and selects at least 16 of them for the next stage.
3. The annotator manually crops at least 8 images from the previous stage, giving at least 8 ground truth windows with good compositions for the assigned subject. Although there may be a few images where it is not easy to identify 8 good cropped images due to a poor foreground or a too-complex background, we only need relatively good compositions to learn differences in quality between cropping windows within each single image. Therefore, we consider selecting 8 best compositions from each image to provide suitable training for our network. Figure 3 shows an example image from SACD, with the 8 ground truth composition windows for this image and a specific subject.
4. In the last stage, the annotator selects the best of the cropped images in the last stage, completing the whole annotation process.

For each image, the whole process produces at least 108 cropping windows $B = \{b_i | 0 \leqslant i < n\}$, where n is the number of cropping windows. Each image contains at least 8 good composition cropping windows B_{good} labeled by annotators, and $B_{\text{good}} = \{g_i | 0 \leqslant i < m\}$, where m is the number of high-quality composition cropping windows, and g_i means b_{g_i} is a good cropping window labeled by annotators. We treat these good cropping windows as ground truth. This process also produces at least 1799 ranked pairs $C = \{(c_{i,1}, c_{i,2}) | 0 \leqslant i < k\}$, where $(c_{i,1}, c_{i,2})$ means $B_{c_{i,1}}$ has better composition than $B_{c_{i,2}}$, and k is the number of ranked pairs.

SACD contains 2777 images, over 24,000 cropping windows labeled by annotators, and over 5.2 million ranked pairs. We divide the SACD dataset into training, validation, and testing sets in the ratios of 8 : 1 : 1.

We compare the annotation process of CPC, GAICD, SACD, and a simple strategy in which annotators directly mark the window with the best composition. Outcomes are reported in Table 1. In order to compare the advantages and disadvantages of each annotation process, we count the number of bounding boxes generated by each annotation process for an image (Boxes), the number of ranked pairs (Pairs), and the number of operations required to annotate an image (Ops). We count one mouse click as one operation and record the number of ranked pairs that are produced by one operation on average (Avg. pairs). Our annotation process produces the most bounding boxes for each image, and each operation produces more ranked pairs on average. In addition, the annotation processes of CPC and GAICD only allow comparison between a few given boxes, but can not directly mark boxes (MB), so lack flexibility and may cause the boxes generated by the annotation process to fail to include

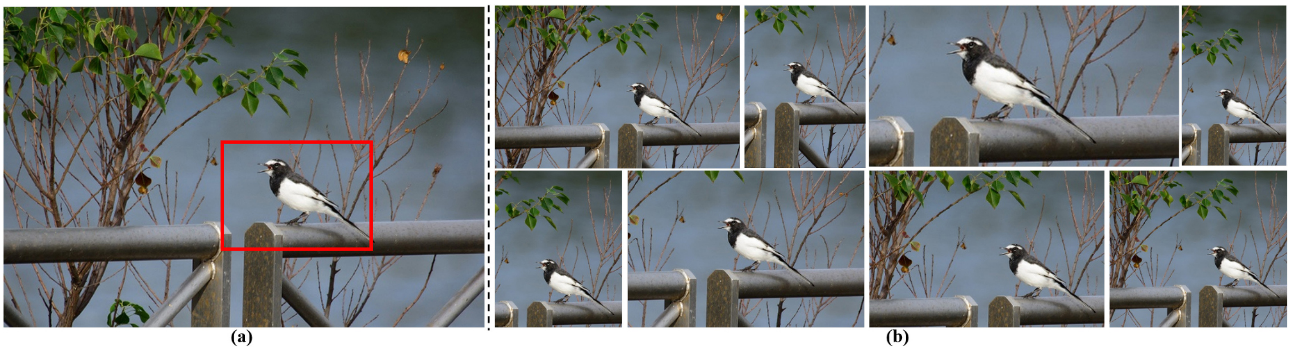


Fig. 3 Example ground truth windows in SACD. (a) An original image; the red box represents the assigned subject. (b) Eight ground truth composition windows for this image, containing the assigned subject.

Table 1 Comparison of annotation processes

Dataset	Boxes	Pairs	Ops	Avg. pairs	MB	CR
CPC	24	63	3	21	N	—
GAICD	87	2958	108	27.4	N	—
Mark	1	0	2	0	Y	256 × 256
SACD	108	2643	53	49.9	Y	51 × 51

the best composition in the image.

Directly marking the best composition bounding box in an image is hard for the annotator, because the candidate range (CR) is the entire image. In the SACD annotation process, for a 256 × 256 image, annotators only need to select the upper left and lower right corners of the bounding box from two 51 × 51 patches respectively when labeling the boxes: the number of candidates is much fewer than for previous methods. Too many candidates will make it more likely that annotators will miss the best composition bounding box. Fewer candidates can provide more objective and accurate annotation results, leading to an easier annotation process and shorter annotation time.

In summary, our 4-stage annotation process can significantly reduce annotation errors and annotation costs while generating a higher number of ranked pairs.

4 SAC-Net

4.1 Outline

Our *subject-aware image composition recommendation network* (SAC-Net) is built upon the basic structure of Faster R-CNN [7]. As Fig. 4 shows, SAC-Net is a two-stage network, where we use candidate anchors and perform ROIalign operations as in Faster R-CNN. However, the composition task differs from the object detection task. For example, each object has only one ground-truth bounding box, but one image may have multiple cropping windows with good compositions. Furthermore, composition datasets are annotated by ranking or pair-wisely comparison of different cropping windows for an image, which differs from object detection datasets. To enable the network to predict accurate cropping windows and adapt to the available training data, we made

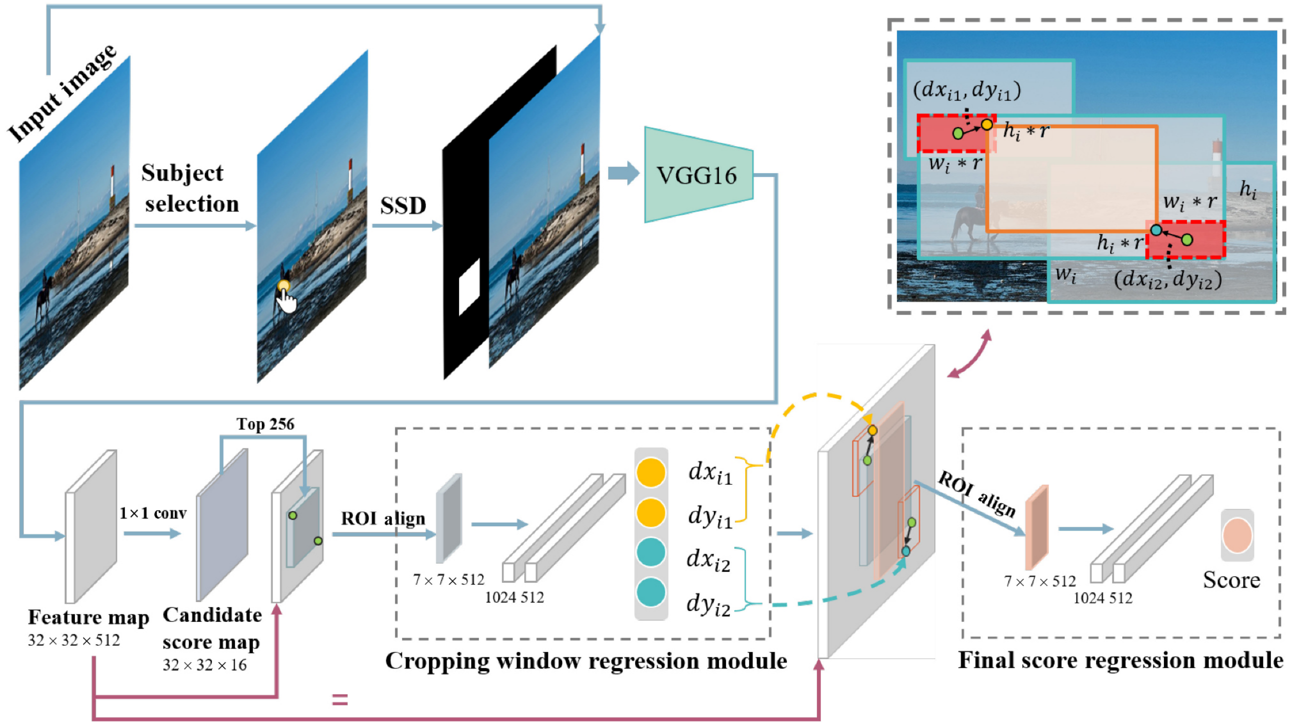


Fig. 4 Inference pipeline of SAC-Net. For an input image, SSD is used to obtain the subject mask after the user selects the subject. The concatenation of the mask and the image are then fed into SAC-Net. SAC-Net has three sub-modules. The *candidate score map regression module* extracts a feature map and convolves it to obtain a candidate score map, in which the top 256 anchors with highest score are selected and ROIaligned. The *cropping window regression module* obtains fine-tuning offsets for the 256 anchor boxes, which are ROIaligned again with fine-tuned boxes on the feature map. The *final score regression module* computes the final composition score of each fine-tuned box. Top right: example of fine-tuning an anchor box.

several improvements to the network structure and proposed new learning schemes. More particularly, a new score continuity loss is used to ensure more reasonable aesthetic score regression.

In this section, we first introduce the proposed modules and the SAC-Net training process using the SACD dataset which enable it to recommend cropping windows, then introduce the composition score continuity loss tailored to the composition task to improve the performance of SAC-Net, and finally introduce how SAC-Net works during inferencing to predict the best cropping window for a specified subject.

4.2 Training SAC-Net with SACD

The SAC-Net training process is shown in Fig. 5. It has a backbone for feature extraction and three sub-modules to allow effective training for the cropping recommendation task. The candidate score map regression module is used to generate candidate composition cropping windows. The cropping window regression module is used to fine-tune candidate composition cropping windows. The final score regression module is used to regress the composition score of each fine-tuned cropping window. We now consider how each module is trained. In each iteration, the parameters of the backbone and the three sub-modules are all updated to minimize the sum of losses that is introduced below.

Backbone. Our backbone for feature extraction feeds the concatenation of the image and the subject mask into VGG16. In order to get a more detailed feature map, we upsample the feature maps of the last two stages of VGG16 to 32×32, and concatenate the first 256 channels from each of the two stages into a $32 \times 32 \times 512$ tensor as our feature map f, which is used as the input to the following three sub-modules.

Candidate score map regression. We use $32 \times 32 \times 16$ anchors to represent possible cropping windows. Each anchor $a_{i,j,k}$ $(0 \leqslant i, j \leqslant 31, 0 \leqslant k \leqslant 15)$ represents a rectangular window centred at pixel $(8i+4, 8j+4)$, of size $w_k h_k$, where w_k and h_k are calculated by the k-means [55] clustering method. The candidate score map regression module aims to give higher scores to anchors with better composition. We use the ranked pair set C from the SACD dataset to learn a score p_i for each anchor a_i. However, for most anchors, there is no annotated cropping window aligned to them. We thus use the annotated window with the largest intersection-over-union (IoU) with the anchor. We set $q_i = t$ if a_t has the largest IoU for all anchors with b_i. We use *candidate score distance loss* L_{CSD} following Ref. [2] to train this module:

$$L_{\mathrm{CSD}} = \frac{1}{k} \sum_{i=1}^{k} \max(1 + p_{q_{c_{i,2}}} - p_{q_{c_{i,1}}}, 0) \qquad (1)$$

where k is the number of ranked pairs, $(c_{i,1}, c_{i,2})$ is a ranked pair from ranked pair set C, and $q_{c_{i,1}}$ is a cropping window with better composition than $q_{c_{i,2}}$. We use this loss to supervise the prediction of the candidate score map with a size of $32 \times 32 \times 16$.

Cropping window regression. We use a set

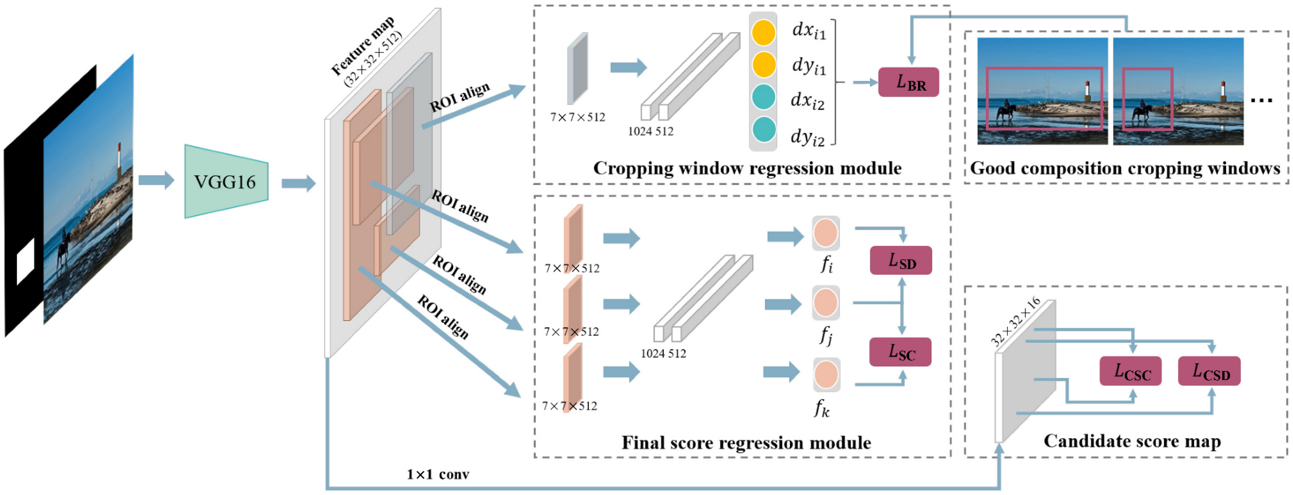


Fig. 5 SAC-Net training pipeline. During training, the candidate score map, cropping window regression module, and final score regression module calculate different losses. The candidate score map and final score regression module are supervised by ranked pairs; the cropping window regression module is supervised by good composition cropping windows. Losses of each module not only affect that module, but also the backbone.

of good cropping windows B_{good} from the SACD dataset as ground truth to supervise the training of the cropping window regression module. We feed each anchor a_i into it and get a refined cropping window W_i. If a good cropping window b_{g_j} is in the regression range of a_i, we mark it as $b_{g_j} \in \Omega(a_i)$. Its refined window W_i should be close to the box b_{g_j}. Here, we design a *box regression loss* to penalize the distance between their corner points.

$$L_{\text{BR}} = \frac{1}{M} \sum_{j=1}^{m} \sum_{i|b_{g_j} \in \Omega(a_i)} |W_i - b_{g_j}| \tag{2}$$

where M is the number of (i, j) satisfying $b_{g_j} \in \Omega(a_i)$, and $|W_i - b_{g_j}|$ represents the L1 distance between the corners of the two boxes. The regression range of an anchor is given in detail in the Electronic Supplementary Materials (ESM).

Final score regression. We use cropping window set B and ranked pair set C from the SACD dataset to supervise the training of the final score regression module in a pair-wise manner. We use each cropping window in B to ROIalign the feature map and send them to the final score regression module to get their scores. Denoting the score of b_i as f_i, we define the *score distance loss* L_{SD} as Eq. (3):

$$L_{\text{SD}} = \frac{1}{k} \sum_{i=1}^{k} \max(1 + f_{c_{i,2}} - f_{c_{i,1}}, 0) \tag{3}$$

Combining the above losses to train SAC-Net allows SAC-Net to learn how to generate final cropping windows and evaluate their compositions:

$$L_{\text{com}} = \lambda_{\text{SD}} L_{\text{SD}} + \lambda_{\text{CSD}} L_{\text{CSD}} + \lambda_{\text{BR}} L_{\text{BR}} \tag{4}$$

where λ_{SD}, λ_{CSD}, and λ_{BR} weight each loss.

4.3 Score continuity loss

We expect that the composition scores of cropping windows should not differ much when their positions and sizes are similar. Therefore, we propose a score continuity loss to encourage smoothness of the predicted scores in the candidate score map regression module and the final score regression module. This loss requires that the larger the IoU of two cropping windows, the closer their composition scores should be.

We now explain how we apply score continuity losses to the two modules. For the candidate score map regression module, we randomly generate s pairs of anchors $\{(a_{1,0}, a_{1,1}), \ldots, (a_{s,0}, a_{s,1})\}$, and predict their scores $p_{a_{i,0}}$ and $p_{a_{i,1}}$. We define the *candidate score continuity loss* L_{CSC} as Eq. (5):

$$L_{\text{CSC}} = \frac{1}{s} \sum_{i=1}^{s} W(\text{IoU}(a_{i,0}, a_{i,1}))(p_{a_{i,0}} - p_{a_{i,1}})^2 \tag{5}$$

where $W(x) = \mathrm{e}^{-(x-1)^2/(2\sigma)}$, with $\sigma = 0.05$. It penalizes large differences between similar anchors.

For the final score regression module, we use a similar idea as for the candidate score map regression module to improve the smoothness of the results. We randomly generate s ranked pairs of cropping windows $\{(u_{1,0}, u_{1,1}), \ldots (u_{s,0}, u_{s,1})\}$, and predict their scores $f_{u_{i,0}}$ and $f_{u_{i,1}}$. If a pair of cropping windows are close to each other, we use the *final score continuity loss* L_{SC} to ensure neighboring anchors have similar scores; it is defined as

$$L_{\text{SC}} = \frac{1}{s} \sum_{i=1}^{s} W(\text{IoU}(u_{i,0}, u_{i,1}))(f_{u_{i,0}} - f_{u_{i,1}})^2 \tag{6}$$

where $W(x) = \mathrm{e}^{-(x-1)^2/(2\sigma)}$. We normally set $\sigma = 0.05$.

Adding the score continuity loss to Eq. (4), our final overall loss function is as Eq. (7):

$$\begin{aligned} L =& \lambda_{\text{SD}} L_{\text{SD}} + \lambda_{\text{CSD}} L_{\text{CSD}} + \lambda_{\text{BR}} L_{\text{BR}} + \\ & \lambda_{\text{SC}} L_{\text{SC}} + \lambda_{\text{CSC}} L_{\text{CSC}} \end{aligned} \tag{7}$$

where λ_{SC} and λ_{CSC} weight each score continuity loss.

4.4 Inferencing

We now introduce how the inferencing stage of SAC-Net works to recommend cropping windows for a specific subject. The workflow is shown in Fig. 4; the ESM gives a more detailed description. Given an input image, the user selects a subject that needs to be composed with a single click. We then use single shot detection (SSD) to get the bounding box mask of the selected subject. We resize the concatenation of the subject mask and the input image to a 256 × 256 × 4 tensor for input to SAC-Net.

Candidate score map regression module. SAC-Net first obtains the feature map f from the input through the backbone. We use 32 × 32 × 16 anchors to represent the possible cropping windows, and feed f to a 1 × 1 convolutional layer to get a 32 × 32 × 16 candidate score map. The 256 anchor boxes with highest scores are fed into the modules of the second stage.

Cropping window regression. These anchor boxes need to be refined according to the image content to get the final cropping windows, since the anchors only provide coarse cropping parameters. We use the anchor box to ROIalign the feature map f and get a $7 \times 7 \times 512$ tensor, and then use a 2-layer full connection to regress the fine-tuned window as shown in Fig. 4.

Final score regression module. Since the candidate score map only provides the aesthetic score of the coarse-level cropping windows, we further predict accurate scores for the regressed windows. Most two-stage detection frameworks such as Ref. [7] regress the offset and the category label of the bounding box simultaneously, since a small change in bounding box position or size does not change the semantic information for the box. However, in the composition task, minor changes to cropping windows affect their composition quality. Therefore, SAC-Net first regresses the fine-tuned window, and then ROIaligns the feature map with a fine-tuned window to regress the final score. We finally sort all the refined boxes according to their final composition scores, and take the box with the highest score as the output cropping window.

5 Experiments

5.1 Overview

We evaluated SAC-Net on the test set of SACD dataset and conducted a user study to validate the reliability of our method. The results show that our model outperforms other existing methods by a large margin. We also conducted comparisons using existing datasets [51] with satisfactory results, indicating that SAC-Net also works well for general composition tasks without awareness of the main subject. In addition, we evaluated our score continuity, showing that it makes composition score prediction more stable and continuous when moving cropping windows, with potential benefits in real-time composition guidance applications.

5.2 Implementation details

In order to avoid cropping windows from exceeding image boundaries, we limited windows predicted by the network to the image size during inferencing. During training, loss weights were set to $\lambda_{\mathrm{SD}} = 1$, $\lambda_{\mathrm{CSD}} = 1$, $\lambda_{\mathrm{BR}} = 10$, $\lambda_{\mathrm{SC}} = 30$, and $\lambda_{\mathrm{CSC}} = 30$. During inferencing, we use single shot detection (SSD) [56] trained on the Pascal visual object classes 2007 (VOC2007) dataset [57] to generate the subject window mask; it runs 45.5 frames per second (fps) so satisfies the requirements of real-time applications. In all experiments, our model is initialized with Faster R-CNN [7] trained on the VOC2007 dataset and uses a stochastic gradient descent (SGD) solver with a batch size of 1. The learning rate is initially set to 0.0001 and multiplied by 0.1 every 200k iterations. We trained our model on an NVIDIA GTX 1080Ti GPU.

5.3 Evaluation metrics

We use intersection-over-union (IoU) and boundary displacement (Disp.) as our evaluation metrics, following previous works [2, 6]. Each image in the SACD dataset has at least 8 good cropping windows labeled by professional artists. When testing, we calculate the IoU between the best cropping window predicted by each method and each ground truth window. We use the ground truth window with the highest IoU with the output cropping window to generate the final IoU and Disp. values. For all images in the SACD dataset test set, we use the mean of IoU and Disp. values to evaluate the performance of each method. We also tested all methods on the FLMS dataset [51] following the above strategy like previous methods [2, 3].

Other previous works use SRCC and $\mathrm{ACC}_{K/N}$ to evaluate the prediction results [3]. However, they require that each image has been annotated with the composition scores of pre-set cropping windows in their training data to allow the network to learn to generate scores for each pre-set window. The annotation pipeline of SACD does not include a stage where users directly rate pre-set windows, nor does our network generate final scores for pre-set windows. Therefore, we do not use SRCC and $\mathrm{ACC}_{K/N}$ to evaluate the output of our networks.

5.4 Quantitative results

5.4.1 Results on the SACD dataset

Using the SACD dataset test set, we compared the following methods which have released their model and code: A2-RL [39], VFN [1], VEN [2], VPN [2], LVRN [4], and GAIC [3]. Since our model uses additional subject window mask information, for a fair comparison, we perform a post-processing step on these methods to discard their results that do not include 50% of the area of the specified subject. SAC-Net builds upon Faster R-CNN, a method specifically designed for object detection. We have made several key changes to the network for the composition task. To demonstrate the effectiveness of these changes, we compare our model to the baseline model, Faster R-CNN, on the image cropping task. We train the baseline model by inputting an image and its subject mask, and using bounding boxes of good cropping windows to supervise the output. The highest score in the output bounding boxes is used as its result. We denote this model as FRCNN-m.

Results are shown in Table 2. Some outputs of different methods are shown in Fig. 6. Our method achieves the best performance in the subject-aware composition task. Our model can generate cropping windows with good compositions which are more prominent and suitable for different types of subjects. It benefits from our proposed SACD dataset, novel network structure, and dedicated learning paradigms for the subject-aware composition task.

Table 2 Results on the SACD dataset test set

Method	IoU(↑)	Disp.(↓)
A2-RL [39]	0.6674	0.0887
VFN [1]	0.6690	0.0887
VPN [2]	0.7036	0.0699
VEN [2]	0.6911	0.0765
LVRN [4]	0.6962	0.0765
GAIC [3]	0.7124	0.0696
FRCNN-m [7]	0.7306	0.0587
SAC-Net (ours)	**0.7665**	**0.0491**

In addition, our network design considers real-world usage of the cropping recommendation method, e.g.,

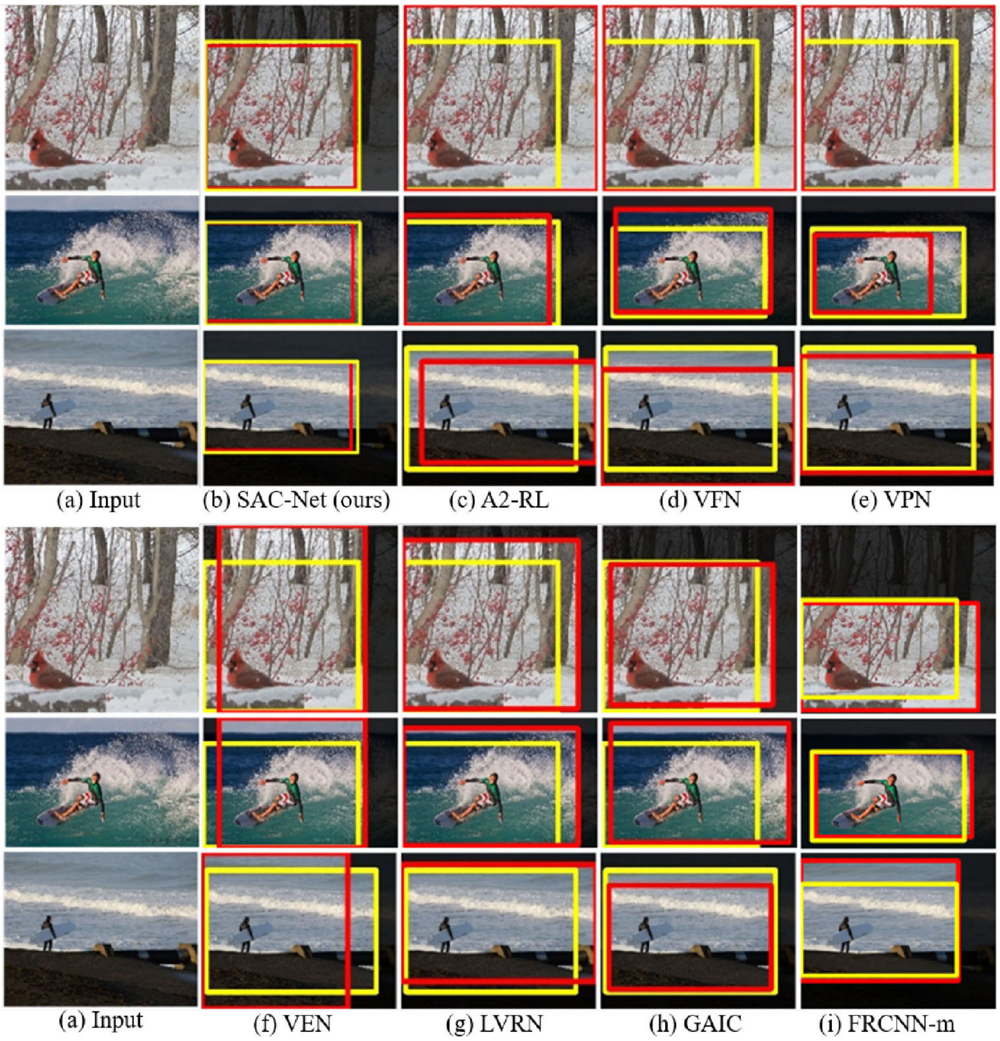


Fig. 6 Results using the SACD dataset test set. We compare our method to 7 other methods (c)–(i). Red box: best cropping window for a method. Yellow box: one of the 8 ground truth windows, with highest IoU to the red box.

automatic composition recommendation in the viewfinder of a digital camera or smart phone. The two score regression sub-modules are designed to obtain similar results for similar cropping windows, in order to provide better spatial stability of predicted scores when providing guidance for users pursuing good compositions. Figure 7 illustrates the score continuity test result. We fix a cropping window in the center of one image, and move the window in different directions. We plot predicted score curves for the moving cropping window using our method, VFV, VEN, GAIC, and LVRN; VPN and A2-RL are excluded because they cannot predict scores of given cropping boxes. We calculate the average discrete curvature of each curve, as report them in Table 3. The result shows that our method gives more stable scores when the camera is moving, so our method provides more stable cropping windows. Please see the video in the ESM for a visual comparison.

5.4.2 User study on single object images

To further evaluate the effectiveness of our proposed method, we conducted a user study with 76 participants aged from 17 to 35, among whom were 21 professional photographers and 55 novice users. We randomly selected 30 pictures from the SACD dataset test set, and generated the best cropping window using 7 different methods. We then showed the original photo, its subject, and 7 sub-images cropped by different methods for each input image to participants. Participants were asked to select the sub-image with the highest aesthetic quality with respect to the specified subject. Considering that different methods may generate similar results, we allow participants to choose from 1 to 3 sub-images. We compared our method to VPN, GAIC, VEN, A2-RL, LVRN, and VFN. We obtained a total of 76 valid results, and calculated the average votes for each method, which are plotted in Fig. 8. Most professional and novice users prefer the results cropped by our method when considering the given subject. It shows the effectiveness of our network and the reliability of our dataset. More details can be found in the ESM.

Table 3 Mean discrete curvature of plots in Fig. 7

Method	X dir.(↓)	Y dir.(↓)	XY dir.(↓)
VFN [1]	0.9056	0.4372	1.2840
VEN [2]	0.1170	0.1668	0.2807
GAIC [3]	0.0288	0.0386	0.0432
LRVN [4]	0.0041	0.0036	0.0156
SAC-Net	**0.0027**	**0.0031**	**0.0093**

5.4.3 User study on multiple objects images

When there are multiple objects in a scene, users may expect different compositions when focusing on different main subjects. To validate the above hypothesis and test the capability of SAC-Net when predicting compositions for multi-object images, we conducted a second user study. We randomly selected 6 images containing at least two objects from both CPC and SACD test sets to give a total of 12 images with 24 subjects. We then generated the 2 cropping windows for each image with 2 different specified main subjects, using 7 different methods (SAC-Net, VPN, GAIC, VEN, A2-RL, LVRN, and VFN). For methods other than SAC-Net we performed a post-processing step to discard results that did not include 50% of the area of the specified subject. We recruited 37

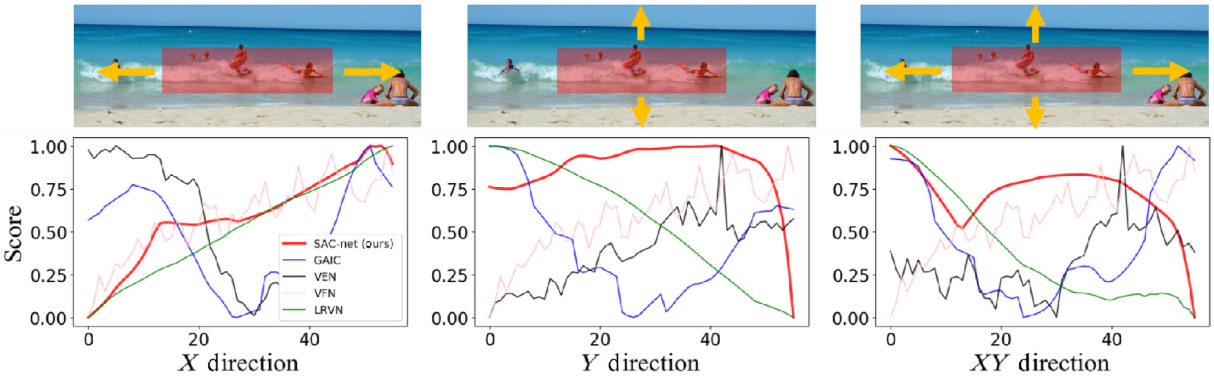


Fig. 7 Score continuity test. We fix a cropping window in the center of one image, expand the window in different directions, and plot the score curve predicted by five different methods. x-axis: number of pixels the cropping window has expanded in the corresponding direction. y-axis: score of the current windows.

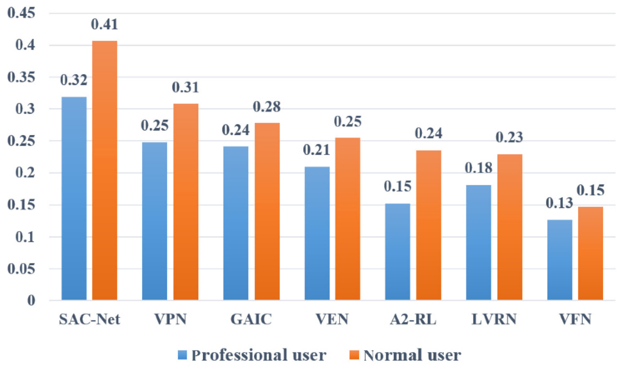


Fig. 8 User study results for single-object images using the SACD dataset test set.

participants aged from 18 to 41, among whom were 12 professional photographers and 25 novice users. We showed 24 sets of images to all participants; each set contained 7 images cropped by 7 different methods for a specified subject. Participants were asked to select the cropped image with the best aesthetic quality with respect to the specified subject from each set of cropped images. Participants were allowed to choose at most 3 images from each set. Average votes for each method are shown in Fig. 9.

The results demonstrate that our method has a greater advantage for scenes with multiple objects than for scenes with a single object (votes shown in Fig. 8). As shown in Fig. 10, methods for predicting general compositions have difficulty in producing a good composition targeting a specific subject, even if the generated composition contains the specified subject. SAC-Net takes the position and size of the main subject into account when generating the final cropping window, which results in good aesthetic quality even in such complex scenes. More details of this user study can be found in the ESM.

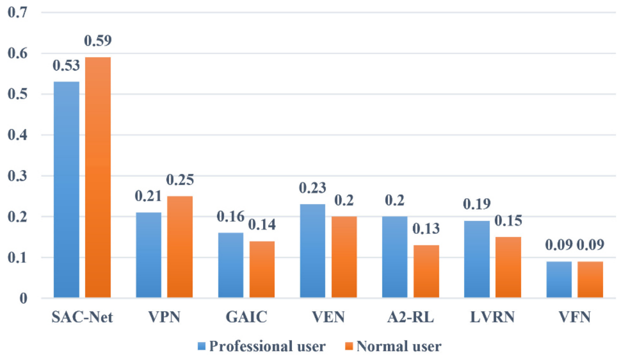


Fig. 9 User study result on multiple-object images using the SACD and CPC dataset test sets.

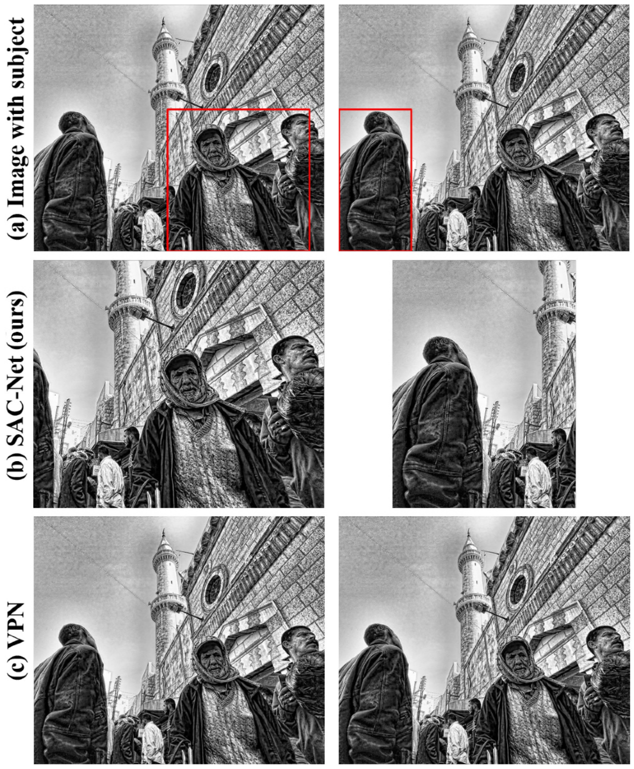


Fig. 10 Sample images used in the second user study. (a) Input image with multiple objects; red box indicates the specified subject. (b) Cropping results of SAC-Net. (c) Cropping results of VPN, an example of methods predicting general compositions.

5.4.4 Results on the FLMS dataset

In order to verify the generality of our proposed network architecture for composition recommendation, we also compared our method to the others on the FLMS dataset [51], which has 500 images and is widely used in the general composition task. Since FLMS is a test set without subjects, for a fair comparison, we used the CPC [2] dataset to train our model, removed the subject mask channel from the input, and weights of all losses during training were set to $\lambda_{\text{SD}} = 1$, $\lambda_{\text{CSD}} = 1$, $\lambda_{\text{BR}} = 10$, $\lambda_{\text{SC}} = 3$, and $\lambda_{\text{CSC}} = 3$. We ran the k-means algorithm to get anchor sizes as described in Section 4.2. We compared the following methods to ours: A2-RL [39], VFN [1], VEN [2], VPN [2], LVRN [4], GAIC [3], ASM-Net [6], and Faster R-CNN [7] trained on CPC, denoted as FRCNN-m. The results are shown in Table 4. Although our method is designed for subject-aware composition, it can still achieve state-of-the-art performance on the general composition task, showing that our proposed method augments the existing literature.

Table 4 Results on FLMS dataset

Method	IoU(↑)	Disp.(↓)
A2-RL [39]	0.8136	0.0472
VFN [1]	0.7371	0.0615
VEN [2]	0.8126	0.0438
VPN [2]	0.8233	0.0399
LVRN [4]	0.8436	0.0365
GAIC [3]	0.8025	0.0470
ASM-Net [6]	0.8486	0.0390
FRCNN-m [7]	0.8125	0.0757
SAC-Net (ours)	**0.8551**	**0.0333**

5.4.5 Speed

We further compared the efficiency of our method to A2-RL, VFN, VEN, VPN, LVRN, GAIC, and ASM-Net; results in terms of frames per second are shown in Table 5, indicating that our method is faster than most algorithms. Where it is slightly slower, our method gives much better results. The result for SSD+SAC-Net shows the efficiency of our method with SSD used for preprocessing. Note that even when taking the time for processing user input into consideration, our method can still meet the real-time requirements. In addition, as Fig. 11 shows, our method is able to balance speed and quality to meet different needs by simply adjusting a parameter.

Table 5 Efficiency evaluation

Method	fps(↑)
LVRN [4]	125
GAIC [3]	125
VPN [2]	75
VEN [2]	0.2
A2-RL [39]	4.1
VFN [1]	0.5
SAC-Net (ours)	100
SSD+SAC-Net	31.3

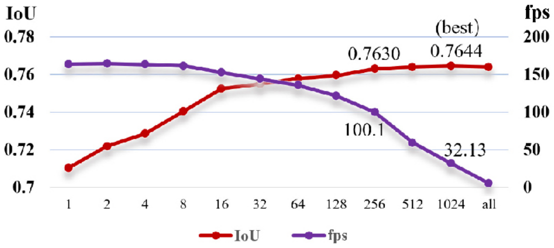


Fig. 11 Varying use of candidate score map. We show IoU and fps values using cropping windows whose scores in the candidate score map are in the top 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, all.

5.4.6 Robustness to subject mask change

SAC-Net obtains the subject information by using SSD to generate a subject bounding box mask. In order to test the robustness of SAC-Net to changes in this mask, we conducted an experiment using inaccurate bounding boxes as input to SAC-Net. We replaced the subject bounding box detected by SSD with a randomly perturbed bounding box using the SACD test set. We experimented with constraints that the IoU of the randomly perturbed bounding boxes and the original bounding boxes should be not less than $x\%$ for varying x. The prediction results are shown in Table 6. When the input mask is relatively accurate, i.e., the IoU of the random box and the original box is greater than 80%, there is almost no effect on the result ($\approx 0.2\%$). When there are moderate errors, i.e., the IoU is greater than 50%, there is a small impact on the result ($\approx 0.6\%$). But if detection completely fails, there is a big impact on the result ($\approx 11.2\%$).

In Fig. 12, we show some of the cropping results of SAC-Net without subject mask input to simulate subject detection failure. We see that our network can still generate acceptable composition results in such cases.

This experiment shows that SAC-Net has good robustness to possible errors in the automatically generated subject masks.

5.5 Ablation experiment

We performed extensive ablation studies on SAC-Net to evaluate the importance of each component, losses, and learning schemes in our method. We tested the model without the candidate score map regression module, cropping window regression module, and final score regression module to provide a better understanding of the contribution of each sub-module. All experiments were conducted on the SACD test set.

Table 6 Subject mask perturbation results

Bounding box IOU	IoU(↑)	Disp.(↓)
0%	0.6550	0.0859
50%	0.7603	0.0509
80%	0.7650	0.0497
90%	0.7659	0.0492
95%	0.7661	0.0492
100%	0.7665	0.0491

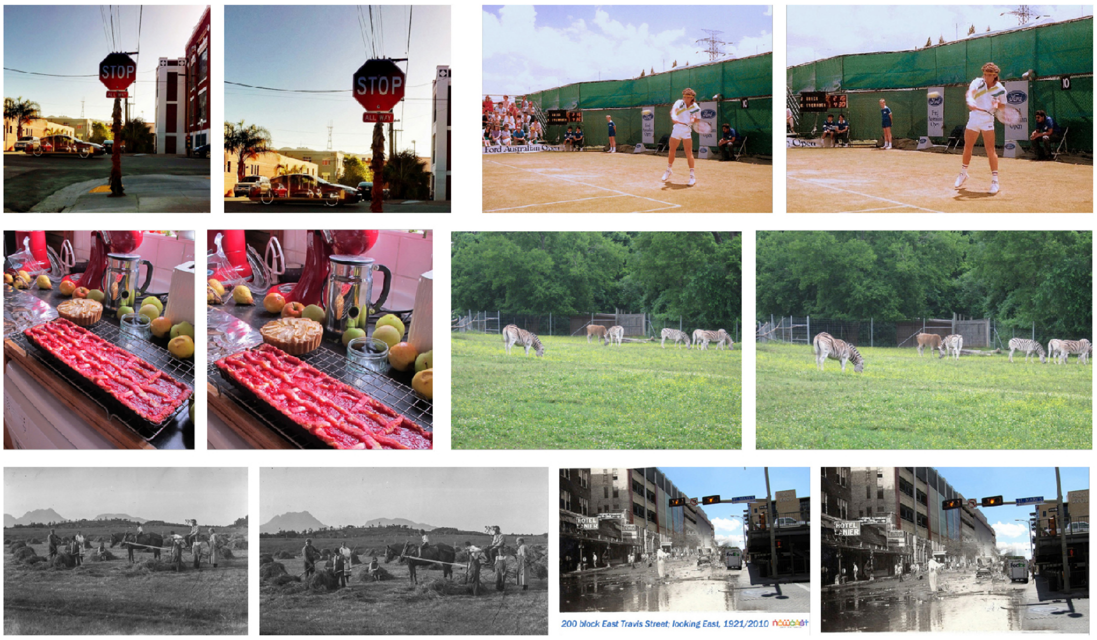


Fig. 12 Cropping results from SAC-Net without subject mask input. In each pair, the input image is on the left and the cropped result on the right.

5.5.1 Score continuity loss

In order to verify effectiveness of the score continuity loss, we conducted an experiment with $\lambda_{\text{SC}} = 0$ and $\lambda_{\text{CSC}} = 0$. This model is denoted as w/o-SC. As reported in Table 7, the score continuity loss not only contributes to the final stability in potential practical applications, but also benefits the cropping window prediction, as it further facilitates score learning by propagating good composition scores to its neighboring anchors and cropping windows.

5.5.2 Candidate score map regression module (CSMRM)

We sent all anchors directly to the cropping window regression module and final score regression module without selecting good candidates using CSMRM. This model is denoted as SAC-Net-all. Results are shown in Fig. 11. Since there is no candidate score to filter out inferior cropping windows, the final IoUs decrease to some extent, and the speed also drops dramatically.

Table 7 Ablation results

Method	IoU(↑)	Disp.(↓)
w/o-CWRM-L	0.7125	0.0736
w/o-CWRM-M	0.7400	0.0582
w/o-CWRM-S	0.5859	0.0714
w/o-FSRM	0.7082	0.0584
w/o-SC	0.7097	0.0622
w/-parallel	0.7239	0.0636
w/o-cluster	0.7320	0.0609
w/o-SM	0.7456	0.0563
w/-LL	0.7512	0.0505
SAC-Net	**0.7665**	**0.0491**

5.5.3 Cropping window regression module (CWRM)

We removed the cropping window regression module, and directly return the largest/mid/smallest window within the anchor regression range to feed the final score regression module, which are denoted w/o-CWRM-L, w/o-CWRM-M, w/o-CWRM-S respectively. The results are reported in Table 7. It can be seen that the window cropping regression module gives SAC-Net better flexibility to adapt to different types of scenes and objects, and hence helps the full model to obtain better cropping windows.

5.5.4 Final score regression module (FSRM)

In order to validate the importance of the score regression module, we directly used scores from the candidate score map after anchor refinement. This model is denoted as w/o-FSRM. As reported in Table 7, there is a large performance drop compared to the full model. It shows that using the estimated final scores of the refined cropping windows can provide much better composition recommendations.

5.5.5 Network structure

In order to enhance the capability of SAC-Net in the composition task, we also make some improvements on the basic network structure compared to Faster R-CNN, including serial score regression, anchor determination with clustering, and feature map expansion. We conducted experiments to verify the effectiveness of these strategies.

Since minor changes to the bounding box in the composition task can affect the composition score of the regressed box, we arranged the box regression module and the score regression module sequentially, instead of running them in parallel as in the common two-stage object detection task [7]. In Table 7, we report the performance when the two modules run in parallel, which is denoted as w/-parallel. The results show that using the two modules sequentially makes better predictions of composition scores of cropping windows.

In order to verify the effectiveness of using k-means to determine anchors, we compared the performance of the network using anchors with manually determined sizes. We make the height and width of each anchor linearly distributed, with w_k={108, 160, 212, 264, 108, 160, 212, 264, 108, 160, 212, 264, 108, 160, 212, 264}, h_k={108, 108, 108, 108, 160, 160, 160, 160, 212, 212, 212, 212, 264, 264, 264, 264}. We denote the model with the above anchor configuration as w/o-cluster. As shown in Table 7, using a clustering method to determine the anchors can evenly distribute potential good composition bounding boxes in the regression range of different anchors, so as to share the regression burden of different anchors, which helps the network to obtain better results.

We also tried different strategies for feature map generation. Instead of using the feature maps from the last two stages of VGG16, we trained a model that directly ROIaligns the final output of VGG16 to a $32 \times 32 \times 512$ tensor as the feature map. This model is denoted as w/-LL. Results are shown in Table 7. The comparison shows that our strategy obtains more detailed information and provides better results by concatenating the last two feature maps of VGG16.

5.5.6 Subject mask

In order to see whether our network learns the composition information with respect to the given subject, we trained our model without the subject mask, denoted w/o-SM. The results reported in Table 7 demonstrate the capability of our network to incorporating subject information from the subject mask channel to give better composition recommendations for specified subjects.

5.5.7 Multiple objects

We mainly target subject-aware composition recommendation, and provide an annotated database and a deep model for composition recommendation for a single subject. Since our network overall learns how to obtain a good cropping window taking into account the content in the mask, some general factors leading to good compositions are learned regardless of the number of masked objects. Thus, we further tested the generalizability of our deep model to cope with photos of multiple objects. We only need to combine the masks of multiple objects as the input mask to achieve the recommended composition for multiple objects. The predicted results of our trained networks using different subject masks are shown in Fig. 14(c), and indicate that our network can generate different cropping windows adapted to different subject sets.

5.6 Further discussion

We also evaluated the effect of the number of candidate cropping windows. After obtaining the candidate score map, we fed all top $2^n (n = 10, \ldots, 1)$ anchors with highest scores to the next sub-modules when testing. The results are shown in Fig. 11. It shows that the candidate score map can speed up the inference process and filter out unreasonable cropping windows for subsequent steps, which is beneficial to the cropping window regression and final score regression modules. The number of candidate cropping windows can be used as a parameter to balance speed and quality. Figure 11 shows that SAC-Net-top-256 as our final model is a good compromise.

In Fig. 13, we show some cases for which SAC-Net failed when testing. Although the composition is not bad for a single subject in those cases, if the network could understand the interaction between the main subject and other objects in the image, it would be more likely to produce higher-quality composition results. We will investigate the effects of interaction between multiple objects in our future research.

We consider that our solution is suitable for applications where it is easy for users to provide the position of the main subject in the current scene. For example, our method could provide a viewfinder assistant for a smart phone camera

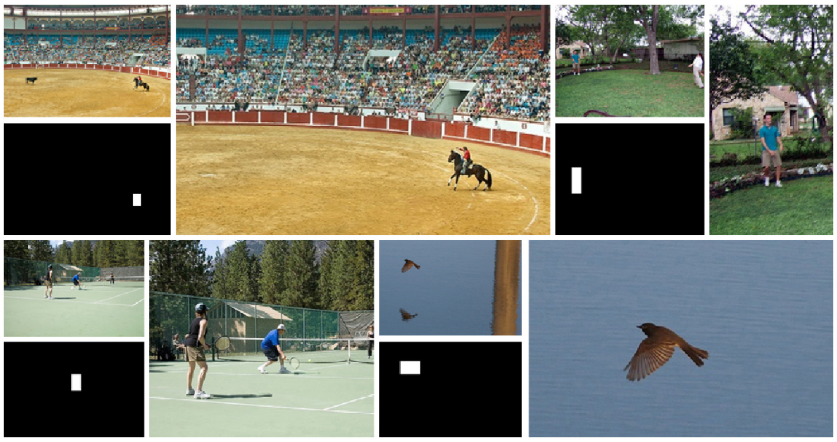


Fig. 13 Failures of SAC-Net on the SACD test set.

application where users can simply tap the subject's position; it could also be used as an auto-cropping tool in image editing software to help novice users to achieve aesthetically pleasing photos as output. Our method can automatically provide the composition for a subject in the above scenarios. The requirement of the existence of a main subject naturally leads to a limitation that landscape pictures with no obvious subject are not handled very well.

Some recommendation results for various situations are shown in Fig. 14 and more results can be seen in the ESM.

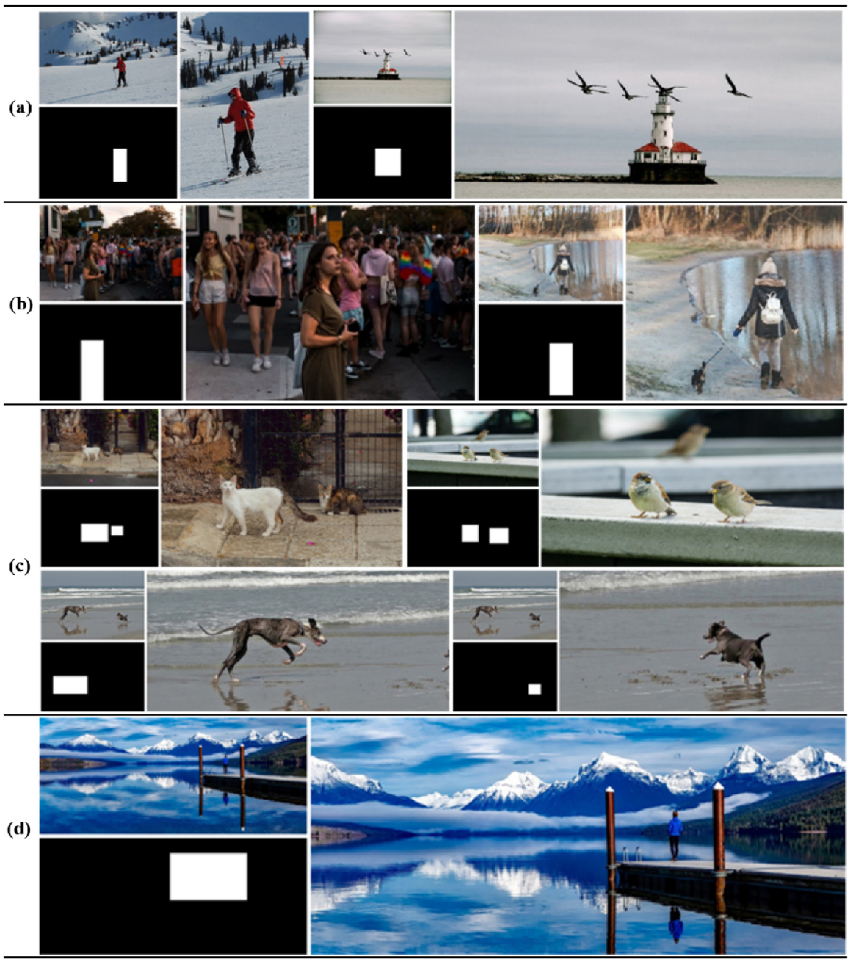


Fig. 14 Further results in different situations. (a, b) Cropping results for a single subject from the SACD dataset and wild images respectively. (c) Results for multiple objects. (d) Results for panoramic images.

6 Conclusions

We have proposed a novel deep model, SAC-Net, and built a new dataset, SACD, for subject-aware composition recommendation. Our model outperforms existing methods on the subject-aware image cropping task, and achieves state-of-the-art performance on the general composition task. Our experiments show that our model is capable of providing good compositions according to the type of subject, and has potential utility in practical applications. Our method is complementary to the literature on automatic cropping for aesthetics. In future, we hope to incorporate semantic relationships between objects, to recommend cropping windows more intelligently. We also intend to use our proposed labeling scheme to collect datasets multiple object composition recommendation and, and hope to develop better deep models to predict crops with good compositions for multiple objects.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (61521002, 62132012) and the Marsden Fund Council managed by the Royal Society of New Zealand (MFP-20-VUW-180).

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Electronic Supplementary Material

Supplementary material is available in the online version of this article at `https://doi.org/10.1007/s41095-021-0263-3`.

References

[1] Chen, Y. L.; Klopp, J.; Sun, M.; Chien, S. Y.; Ma, K. L. Learning to compose with professional photographs on the web. In: Proceedings of the 25th ACM International Conference on Multimedia, 37–45, 2017.

[2] Wei, Z. J.; Zhang, J. M.; Shen, X. H.; Lin, Z.; Mech, R.; Hoai, M.; Samaras, D. Good view hunting: Learning photo composition from dense view pairs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5437–5446, 2018.

[3] Zeng, H.; Li, L. D.; Cao, Z. S.; Zhang, L. Reliable and efficient image cropping: A grid anchor based approach. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5942–5950, 2019.

[4] Lu, W. R.; Xing, X. F.; Cai, B. L.; Xu, X. M. Listwise view ranking for image cropping. *IEEE Access* Vol. 7, 91904–91911, 2019.

[5] Freeman, M. *The Photographer's Eye: Composition and Design for Better Digital Photos*. Focal Press, 2007.

[6] Tu, Y.; Niu, L.; Zhao, W. J.; Cheng, D. W.; Zhang, L. Q. Image cropping with composition and saliency aware aesthetic score map. *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 34, No. 7, 12104–12111, 2020.

[7] Ren, S. Q.; He, K. M.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks In: Proceedings of the 28th International Conference on Neural Information Processing Systems, Vol. 1, 91–99, 2015.

[8] Zhang, L. M.; Song, M. L.; Zhao, Q.; Liu, X.; Bu, J. J.; Chen, C. Probabilistic graphlet transfer for photo cropping. *IEEE Transactions on Image Processing* Vol. 22, No. 2, 802–815, 2013.

[9] Chang, Y. Y.; Chen, H. T. Finding good composition in panoramic scenes. In: Proceedings of the IEEE 12th International Conference on Computer Vision, 2225–2231, 2009.

[10] Nishiyama, M.; Okabe, T.; Sato, Y.; Sato, I. Sensation-based photo cropping. In: Proceedings of the 17th ACM International Conference on Multimedia, 669–672, 2009.

[11] Ke, Y.; Tang, X. O.; Jing, F. The design of high-level features for photo quality assessment. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 419–426, 2006.

[12] Dhar, S.; Ordonez, V.; Berg, T. L. High level describable attributes for predicting aesthetics and interestingness. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, 1657–1664, 2011.

[13] Chen, L. Q.; Xie, X.; Fan, X.; Ma, W. Y.; Zhang, H. J.; Zhou, H. Q. A visual attention model for adapting images on small displays. *Multimedia Systems* Vol. 9, No. 4, 353–364, 2003.

[14] Ge, S. M.; Jin, X.; Ye, Q. T.; Luo, Z.; Li, Q. Image editing by object-aware optimal boundary searching and mixed-domain composition. *Computational Visual Media* Vol. 4, No. 1, 71–82, 2018.

[15] Suh, B.; Ling, H. B.; Bederson, B. B.; Jacobs, D. W. Automatic thumbnail cropping and its effectiveness. In: Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology, 95–104, 2003.

[16] Zhang, F. L.; Wang, M.; Hu, S. M. Aesthetic image enhancement by dependence-aware object recomposition. *IEEE Transactions on Multimedia* Vol. 15, No. 7, 1480–1490, 2013.

[17] Marchesotti, L.; Cifarelli, C.; Csurka, G. A framework for visual saliency detection with applications to image thumbnailing. In: Proceedings of the IEEE 12th International Conference on Computer Vision, 2232–2239, 2009.

[18] Xu, P. F.; Ding, J. Q.; Zhang, H.; Huang, H. Discernible image mosaic with edge-aware adaptive tiles. *Computational Visual Media* Vol. 5, No. 1, 45–58, 2019.

[19] Zhang, S. H.; Zhou, Z. P.; Liu, B.; Dong, X.; Hall, P. What and where: A context-based recommendation system for object insertion. *Computational Visual Media* Vol. 6, No. 1, 79–93, 2020.

[20] Sheng, K. K.; Dong, W. M.; Huang, H. B.; Chai, M. L.; Zhang, Y.; Ma, C. Y.; Hu, B.-G. Learning to assess visual aesthetics of food images. *Computational Visual Media* Vol. 7, No. 1, 139–152, 2021.

[21] Luo, J. Subject content-based intelligent cropping of digital photos. In: Proceedings of the IEEE International Conference on Multimedia and Expo, 2218–2221, 2007.

[22] Stentiford, F. Attention based auto image cropping. In: Proceedings of the 5th International Conference on Computer Vision Systems, 2007.

[23] Santella, A.; Agrawala, M.; DeCarlo, D.; Salesin, D.; Cohen, M. Gaze-based interaction for semi-automatic photo cropping. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 771–780, 2006.

[24] Cheng, B.; Ni, B. B.; Yan, S. C.; Tian, Q. Learning to photograph. In: Proceedings of the 18th ACM International Conference on Multimedia, 291–300, 2010.

[25] Rawat, Y. S.; Kankanhalli, M. S. Context-based photography learning using crowdsourced images and social media. In: Proceedings of the Proceedings of the 22nd ACM International Conference on Multimedia, 217–220, 2014.

[26] Yan, J. Z.; Lin, S.; Kang, S. B.; Tang, X. O. Change-based image cropping with exclusion and compositional features. *International Journal of Computer Vision* Vol. 114, No. 1, 74–87, 2015.

[27] Liang, Y.; Wang, X. T.; Zhang, S. H.; Hu, S. M.; Liu, S. X. PhotoRecomposer: Interactive photo recomposition by cropping. *IEEE Transactions on Visualization and Computer Graphics* Vol. 24, No. 10, 2728–2742, 2018.

[28] Su, H. H.; Chen, T. W.; Kao, C. C.; Hsu, W. H.; Chien, S. Y. Preference-aware view recommendation system for scenic photos based on bag-of-aesthetics-preserving features. *IEEE Transactions on Multimedia* Vol. 14, No. 3, 833–843, 2012.

[29] Yan, J. Z.; Lin, S.; Kang, S. B.; Tang, X. O. Learning the change for automatic image cropping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 971–978, 2013.

[30] Kang, L.; Ye, P.; Li, Y.; Doermann, D. Convolutional neural networks for no-reference image quality assessment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1733–1740, 2014.

[31] Lu, X.; Lin, Z.; Jin, H. L.; Yang, J. C.; Wang, J. Z. RAPID: Rating pictorial aesthetics using deep learning. In: Proceedings of the 22nd ACM International Conference on Multimedia, 457–466, 2014.

[32] Lu, X.; Lin, Z.; Shen, X. H.; Mech, R.; Wang, J. Z. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In: Proceedings of the IEEE International Conference on Computer Vision, 990–998, 2015.

[33] Kong, S., Shen, X., Lin, Z., Mech, R., Fowlkes, C. Photo aesthetics ranking network with attributes and content adaptation. In: *Computer Vision – ECCV 2016. Lecture Notes in Computer Science, Vol. 9905.* Leibe, B.; Matas, J.; Sebe, N.; Welling, M. Eds. Springer Cham, 662–679, 2016.

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

[35] Esmaeili, S. A.; Singh, B.; Davis, L. S. Fast-at: Fast automatic thumbnail generation using deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4178–4186, 2017.

[36] Wang, W. G.; Shen, J. B. Deep cropping via attention box prediction and aesthetics assessment. In: Proceedings of the IEEE International Conference on Computer Vision, 2205–2213, 2017.

[37] Wang, W. G.; Shen, J. B.; Ling, H. B. A deep network solution for attention and aesthetics aware photo cropping. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 41, No. 7, 1531–1544, 2019.

[38] Wang, W. G.; Shen, J. B.; Yu, Y. Z.; Ma, K. L. Stereoscopic thumbnail creation via efficient stereo saliency detection. *IEEE Transactions on Visualization and Computer Graphics* Vol. 23, No. 8, 2014–2027, 2017.

[39] Li, D. B.; Wu, H. K.; Zhang, J. G.; Huang, K. Q. A2-RL: Aesthetics aware reinforcement learning for image cropping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8193–8201, 2018.

[40] Chen, Y. L.; Huang, T. W.; Chang, K. H.; Tsai, Y. C.; Chen, H. T.; Chen, B. Y. Quantitative analysis

of automatic image cropping algorithms: A dataset and comparative study. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 226–234, 2017.

[41] Hosu, V.; Goldlücke, B.; Saupe, D. Effective aesthetics prediction with multi-level spatially pooled features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9367–9375, 2019.

[42] Lu, P.; Zhang, H.; Peng, X. J.; Peng, X. Aesthetic guided deep regression network for image cropping. *Signal Processing*: *Image Communication* Vol. 77, 1–10, 2019.

[43] Lu, P.; Zhang, H.; Peng, X. J.; Jin, X. F. An end-to-end neural network for image cropping by learning composition from aesthetic photos. *arXiv preprint* arXiv:1907.01432, 2019.

[44] Li, X. W.; Li, X. M.; Zhang, G.; Zhang, X. L. Image aesthetic assessment using a saliency symbiosis network. *Journal of Electronic Imaging* Vol. 28, No. 2, 023008, 2019.

[45] Lu, P.; Liu, J. H.; Peng, X. J.; Wang, X. J. Weakly supervised real-time image cropping based on aesthetic distributions. In: Proceedings of the 28th ACM International Conference on Multimedia, 120–128, 2020.

[46] Christensen, C. L.; Vartakavi, A. An experience-based direct generation approach to automatic image cropping. *IEEE Access* Vol. 9, 107600–107610, 2021.

[47] Hong, C. Y.; Du, S. Y.; Xian, K.; Lu, H.; Cao, Z. G.; Zhong, W. C. Composing photos like a photographer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7053–7062, 2021.

[48] Datta, R.; Joshi, D.; Li, J.; Wang, J. Z. Studying aesthetics in photographic images using a computational approach. In: *Computer Vision – ECCV 2006. Lecture Notes in Computer Science, Vol. 3953*. Leonardis, A.; Bischof, H.; Pinz, A. Eds. Springer Berlin Heidelberg, 288–301, 2006.

[49] Luo, W.; Wang, X. G.; Tang, X. O. Content-based photo quality assessment. In: Proceedings of the International Conference on Computer Vision, 2206–2213, 2011.

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

[51] Fang, C.; Lin, Z.; Mech, R.; Shen, X. H. Automatic image cropping using visual composition, boundary simplicity and content preservation models. In: Proceedings of the 22nd ACM International Conference on Multimedia, 1105–1108, 2014.

[52] Chang, H. W.; Yu, F.; Wang, J.; Ashley, D.; Finkelstein, A. Automatic triage for a photo series. *ACM Transactions on Graphics* Vol. 35, No. 4, Article No. 148, 2016.

[53] Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C. L. Microsoft COCO: Common objects in context. In: *Computer Vision – ECCV 2014. Lecture Notes in Computer Science, Vol. 8693.* Fleet, D.; Pajdla, T.; Schiele, B.; Tuytelaars, T. Eds. Springer Cham, 740–755, 2014.

[54] He, K. M.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, 2980–2988, 2017.

[55] MacQueen, J. Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, 281–297, 1967.

[56] DeGroot, M.; Brown, E. SSD: Single shot multibox object detector, in PyTorch. 2018. Available at https://github.com/amdegroot/ssd.pytorch.

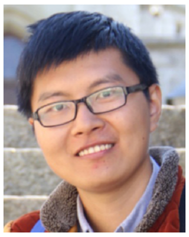
[57] Everingham, M.; van Gool, L.; Williams, C. K. I.; Winn, J.; Zisserman, A. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. 2007. Available at http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

Guo-Ye Yang received his B.S. degree from Tsinghua University, Beijing, China, in 2019, where he is currently pursuing a Ph.D. degree. His research interests include computer graphics, image analysis, and computer vision.

Wen-Yang Zhou received his B.S. degree from Jilin University in 2019. He is currently pursuing a Ph.D. degree in Tsinghua University. His research interests include computer graphics, image analysis, and computer vision.

Yun Cai received her B.S. degree from the Communication University of China in 2018. She is currently pursuing a master degree in Tsinghua University. Her research interests include computer graphics, computer vision, and human–computer interaction.

Song-Hai Zhang received his Ph.D. degree from Tsinghua University in 2007. He is currently an associate professor in the Department of Computer Science and Technology, Tsinghua University. His research interests include image and video processing and geometric computing.



Fang-Lue Zhang is currently a lecturer at the Victoria University of Wellington. He received his Ph.D. degree from Tsinghua University in 2015. His research interests include image and video editing, computer vision, and computer graphics. He received a Victoria Early-Career Award in 2019 and a Marsden Fast-Start Grant from the Royal Society of New Zealand in 2020. He is a member of the IEEE Central New Zealand Sector Committee.