Expert Systems With Applications 186 (2021) 115852

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A grid anchor based cropping approach exploiting image aesthetics, geometric composition, and semantics

Luigi Celona *, Gianluigi Ciocca, Paolo Napoletano

Department of Informatics, Systems and Communication, University of Milano-Bicocca, viale Sarca, 336, 20126, Milano, Italy

ARTICLE INFO

Keywords:
Image cropping
Image aesthetics
Image composition
Semantic content
Deep learning

ABSTRACT

Image cropping aims at the selection of the relevant part of an image maximizing its aesthetic quality and composition. The part of the image that needs to be removed is highly dependent on user preferences and can be related to image aesthetics, composition, informativeness, or other criteria. Since the concept of the perfect crop does not exist, but there are several cropping possibilities, recent cropping algorithms are trained to rank a set of crop candidates based on their compositional quality. To this end, several benchmark databases have been released that provide for each image a series of human-annotated crop candidates with corresponding scores. Many of the image cropping methods rely on a single criterion to define the best crop or crops in an image. However, a single criterion misses the complexity of human opinions which can differ in personal preferences and backgrounds. Motivated by this, we formulate the cropping problem as a ranking problem of candidate crop regions using a grid anchor based approach and multiple criteria. To evaluate the goodness of a crop region, we design a cropping method by combining three efficient and lightweight neural networks specifically designed to evaluate the quality of a crop in terms of aesthetics, composition, and semantics. Our results on standard datasets show that using more criteria yields better crops than state-of-the-art approaches. This result is also confirmed by a subjective study on user preferences that involved a panel of users.

1. Introduction

The aims of image cropping is to re-framing an image by excluding portions of it. Depending on the task, the re-framing is intended to exclude unwanted or non-informative regions, and, in general, to have an image with a better overall photographic composition. Automatic image cropping can be leveraged in different applications comprising digital photography, data visualization, and computer vision.

Similar to image cropping are image thumbnailing, and image retargeting (Cho et al., 2017; Esmaeili et al., 2017). Cropping, thumbnailing, and retargeting are three similar approaches to image re-framing with subtle differences. Image thumbnailing aims at defining a region in the image to be used as a preview of the whole image (Esmaeili et al., 2017; Wang et al., 2016). Thus the thumbnail image should represent the original contents as best as it could. In general, no photographic or compositional constraints are required. Image retargeting is a more complex resizing strategy where the image is *restructured* in order to fit as much information as possible in a target region (Cho et al., 2017). Different from the previous two image re-framing strategies, image cropping often leverage photographic, and compositional rules in order to select the best region in the image. Moreover, depending on the application, the cropped region is usually not constrained with respect to the resolution or aspect ratio (see Fig. 1 for a qualitative comparison among the three different approaches). Many early works in image cropping relied on handcrafted features based on cues borrowed from human's perception or photographic rules (Liu et al., 2010; Nishiyama et al., 2009). More recently, with the development of efficient and effective deep learning techniques, there are many works that exploits them in the context of image cropping using convolutional neural networks.

Most of the recent works in image cropping can be categorized according to the underlying strategy and the cues used in determining the best crop or crops. With respect to the strategy, there are methods that try to find a single best crop in the image (Fang et al., 2014; Li et al., 2018). These methods usually exploits optimization techniques to maximize a given target function that embeds criteria defining what an ideal crop should be (see Fig. 2(a)).

Other approaches use a two step strategy (Liu et al., 2010; Wang & Shen, 2017; Wang et al., 2018). In the first step, a number of candidate regions are selected either via sliding windows or using a set of predefined constrained regions (i.e anchor boxes). In the second step, using suitable criteria, the candidate crops are ranked. The problem thus

* Corresponding author.
E-mail addresses: luigi.celona@unimib.it (L. Celona), gianluigi.ciocca@unimib.it (G. Ciocca), paolo.napoletano@unimib.it (P. Napoletano).

https://doi.org/10.1016/j.eswa.2021.115852
Received 9 June 2021; Received in revised form 30 August 2021; Accepted 31 August 2021
Available online 6 September 2021

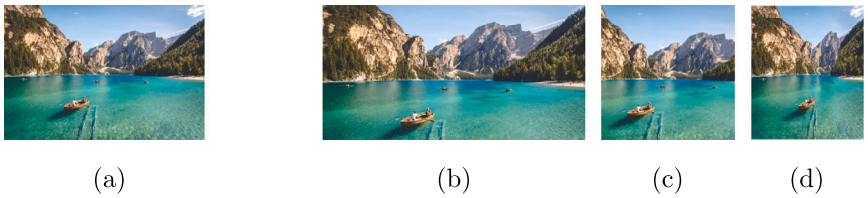

Fig. 1. Comparison between different manipulation techniques for presenting and browsing images. In (a) the original image is presented; (b) shows the result of the proposed cropping method; (c) reports the image produced by the thumbnailing method in Esmaeili et al. (2017); finally, (d) exhibits the original image retargeted to half the width size obtained thanks to Cho et al. (2017).

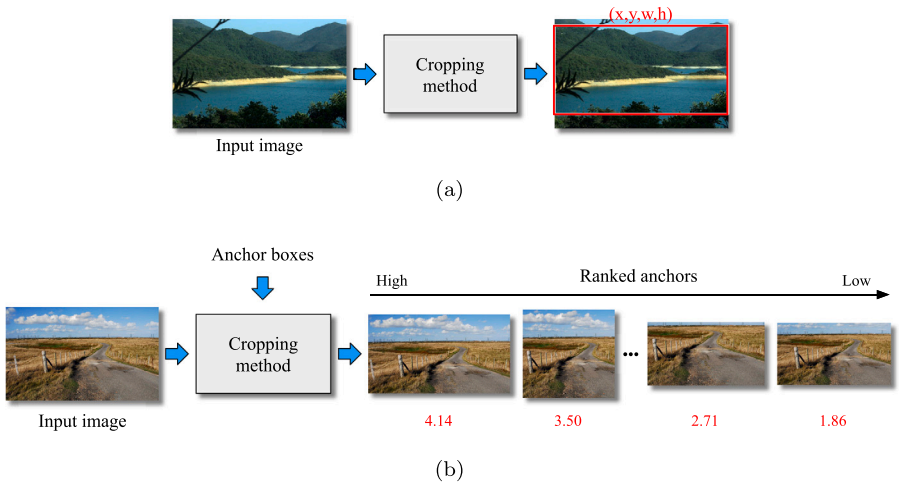

Fig. 2. Cropping paradigms. (a) The goal of this class of cropping methods is to estimate a set of cropping coordinates (x,y,w,h) given an input image. (b) Given an input image, these cropping methods predict a score list by evaluating each anchor box contained in a set of pre-defined anchor boxes.

is reformulated from finding the best crop to determining the most suitable ranking of a set of crops (see Fig. 2(b)).

Regardless of the cropping strategy exploited, cropping algorithms mostly leverage very few criteria in evaluating the goodness of a crop. Most of the cropping algorithms exploits either some notion of saliency or image aesthetic (Ciocca et al., 2007; Kao et al., 2017; Nishiyama et al., 2009; Stentiford, 2007). Few consider these two characteristics together (Lu et al., 2019b; Wang et al., 2018).

Most of the methods in the literature consider either saliency or aesthetic as the most important quality aspect in a crop. We argue that there are other criteria that can be leveraged in image cropping strategies. Moreover, a good cropping strategy should be defined in terms of different attributes that should be simultaneously exploited and that may contribute at different levels.

For these reason, we propose a cropping method that takes into account this rationale. Starting from the available literature, we designed a cropping method where different cropping criteria can be plugged-in in a seamless manner, and that is able to select more meaningful crops with respect to the standard approaches.

Given an image and a set of candidate crop regions as input, our cropping method estimates a score for each crop of the set to pick the best crop region. This is done by considering an ensemble of cropping models each of which is designed to score the candidate crops in terms of a different criterion. The scores are then aggregated and the best crop region is selected.

With respect to the criteria, we currently developed three cropping models that evaluate the aesthetics, geometric composition, and the semantic contents of the crop. The aesthetics and geometry composition are visual criteria that are useful to select the most visually appealing crop region. These criteria are often exploited in works the literature of cropping algorithms. In this work, we introduce the semantic criterion because we want to preserve in the cropped image as much as possible the main content of the original image. This criterion has not been previously considered in designing cropping strategies. We use these three criteria in our experimentation but the cropping models can be easily changed and extended adding other general or user-specific criteria. Our cropping method is fully based on Deep Convolutional Neural Networks. Each cropping model is based on the MobileNet-v2 (Sandler et al., 2018) as a backbone making it efficient and lightweight.

We evaluate our cropping method on four standard cropping datasets: GAICD (Zeng et al., 2019), ICDB (Yan et al., 2013), FLMS (Fang et al., 2014), and FCDB (Chen et al., 2017a). We quantitatively and qualitatively compared different state-of-the-art approaches on these datasets using several reference metrics to assess the methods from different perspectives: Pearson Linear Correlation Coefficient, Spearman Rank-Order Correlation Coefficient, Return K of Top N accuracy, Rank Weighted Return K of top-N accuracy, Intersection-Over-Union, and Boundary Displacement Error. Since the final result of an image cropping algorithm is judged by users, we also conducted a subjective evaluation of the different cropping strategies. The evaluation is performed via a user preference study where the subjects are asked to choose the preferred results among the displayed ones. The experimental results demonstrate the effectiveness of the proposed method.

The main contributions of this paper are:

1. we define the image cropping problem as a ranking problem of candidate crop regions;
2. we design a new cropping method that rank regions by simultaneously leveraging different quality criteria;
3. we develop the cropping method by combining different, efficient and light-weighted, neural networks each of which is specifically designed to evaluate the candidate crops with respect to a specific criterion;
4. we perform an extensive evaluation using several benchmark datasets and quality metrics.

The remainder of this paper is organized as follows. Section 2 gives an overview of related works with particular focus on supervised methods. Section 3 describes our proposed cropping method. Section 4 illustrates the experimental setup. Section 5 presents the results of the experiments. Section 6 concludes this paper.

2. Related work

In this section we summarize the existing image cropping datasets, and the representative image cropping methods.

2.1. Image cropping datasets

In the literature there are datasets with different peculiarities for the training and evaluation of autocropping methods. Over the years, the cardinality of the proposed datasets has increased, also allowing the design of deep learning methods. In addition, the process of creating and acquiring the annotations has changed.

Yan et al. (2013) proposed the CUHK Image Cropping DataBase (ICDB), the first cropping dataset, which consists of 950 images gathered from the CUHKPQ dataset. Images present seven different content types (*i.e.* animal, architecture, human, landscape, night, plant, and static), and each image was manually cropped by three different professional photographers. Fang et al. (2014) collected a similar cropping dataset, which consists of 500 images crawled from Flickr. Images of the FLMS dataset have been cropped by ten professional photographers on Amazon Mechanical Turk (AMT). The ten annotated bounding boxes only have small overlaps meaning that there is little correlation among the ten photographers' knowledge or preference. Chen et al. (2017a) introduced the Flickr Cropping DataBase (FCDB) annotated using a new strategy that is to compare pairs of crops. This strategy allows to enormously increase the number of annotated crops in fact: for 1743 images collected by Flickr, 34,130 pairs of sub-views on AMT were annotated.

Previous datasets present problems regarding (i) the evaluation criterion, and (ii) the generation of candidate crops (Celona et al., 2019; Wei et al., 2018). Therefore, the new datasets were collected with new protocols for candidate crops selection and annotation to provide more reliable and effective evaluation metrics for cropping images. Wei et al. (2018) constructed the Comparative Photo Composition (CPC) dataset. It is a large dataset of 10,797 images. For each image, 24 candidate crops with four standard aspect-ratios have been pooled among candidates automatically generated by exploiting existing recomposition and cropping algorithms. Annotations were collected using a two-stage annotation protocol, which allows to obtain more than 1 million view pairs on AMT. Following the previous protocol, the same authors introduced the eXPert View dataset (XPView), which consists of 992 images annotated by three experts. Zeng et al. (2019) proposed to reduce the searching space of image cropping by defining crops on image grid anchor rather than dense pixels. Thanks to this formulation they constructed the Grid Anchor based Image Cropping dataset (GAICD). It contains a total of 106,860 annotated candidate crops from 1236 source images. The 19 experienced human annotators were required to rate the candidates at five scores. The Mean Opinion Score (MOS) was finally calculated for each candidate crop as ground-truth quality score.

2.2. Image cropping methods

The existing image cropping methods can be divided, according to the taxonomy reported in Fig. 3, in two main categories: weakly supervised methods and supervised methods.

Weakly supervised methods are not trained on annotated bounding boxes but exploit exogenous knowledge to estimate the most salient sub-region on the basis of attention, aesthetics or a combination of both.

In contrast, supervised methods are trained on annotated bounding boxes and in some cases exploit exogenous knowledge in combination with endogenous knowledge to estimate the best sub-region to crop.

The first category of methods generally performs poorly over the second because it does not learn the nature of the problem directly from the data. The second category on the other hand requires a huge amount of data to avoid poor generalization.

These methods can be further categorized on the basis of the information exploited to select the best crop.

2.2.1. Attention-driven methods

Attention-driven methods select the sub-region that contains the most salient subject or the most informative region of an image.

Single-stage. An image cropping method exploiting a generic and extensible image attention model based on three attributes (region of interest, attention value, and minimal perceptible size) has been proposed in Chen et al. (2003). In (Ciocca et al., 2007), Ciocca et al. presented a self-adaptive image cropping algorithm exploiting both visual and semantic information. This algorithm is capable of building a cropping strategy based on the use of a visual attention model specific for different genres of images. Stentiford (2007) presented a method based on the application of different rules to the region identified by the visual attention to cut out the best image sub-region. An approach involving visual composition, boundary simplicity and content preservation has been presented in Fang et al. (2014). The visual composition component consists of a Support Vector Regressor (SVR) to learn a mapping from composition features (saliency maps of original images obtained by using the Spatial Pyramid of Saliency Map) to composition scores. The boundary simplicity encourage crop boundaries to pass through visually simpler regions, in order to reduce the chance of cutting through objects. Finally, content preservation is obtained by exploiting visual saliency. The goodness of the produced crops have been evaluated on the proposed FLMS dataset. Several practical formulations of the optimum rectangle search problem and a new approach describing the relationship between attention preserving and region cropping have been proposed in Chen et al. (2016).

Two-step. Nishiyama et al. (2009) proposed a novel attention-driven technique for image cropping that uses a quality score to choose the best crop among a set of proposals. Given an image, multiple cropping proposals are produced using Itti et al. (1998) in conjunction with the k-Means clustering algorithm. The best of the proposals is chosen considering several basic techniques for photographic (e.g. no camera shakes and adequate exposure).

Iterative. Recently, an automatic photo composition method based on collaborative deep reinforcement learning (called CDRL-RC) has been presented in Li and Zhang (2019). It follows an iterative process where the agent looks at the current cropping window and then sequentially transforms the image windows within the entire image by performing one of the predefined actions until termination. More specifically, two agents, the agent of emotional attention and the agent of the image context, jointly determine the action to be selected from a set of moving and zooming actions. The reward function considers the crop quality and consists of a weighted IoU that takes into account the emotional attention map.

2.2.2. Aesthetic-driven methods

The aesthetics oriented method aims at maximizing the visual attractiveness of the cropped images. Although the visual aesthetics obeys certain general principles, it is also known to be influenced by subjective factors such as the culture, personal experiences, education level, or even the psychological state.

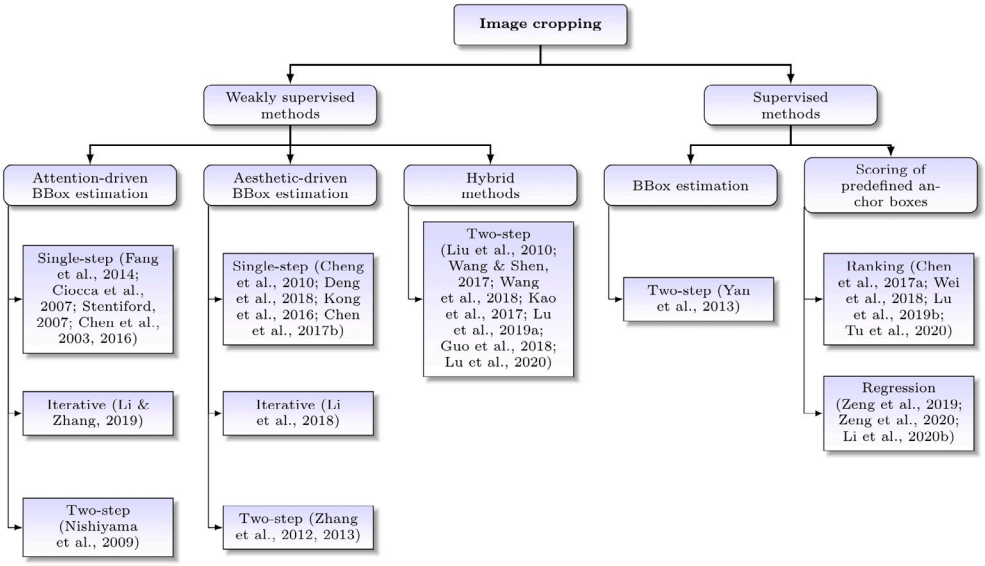

Fig. 3. Taxonomy of image cropping methods.

Single-step. Cheng et al. (2010) proposed a model for encoding professional photographers' knowledge and composition rules, mined from massive crawled professional photos from online sharing website. A photo quality evaluation metric based on a Bayesian algorithm is then used for finding the best image sub-region. EnhanceGAN (Deng et al., 2018) is the first method which adopts adversarial learning to perform color enhancement and image cropping. The latter is performed by using the proposed attentive convolution which outputs 5 feature maps corresponding to the cropping coordinates (x,y,w,h) and a probability map, respectively. Finally Top-K average pooling is used to produce the final crop window.

A deep model has been trained for ranking candidate crops gathered with sliding window strategy by jointly learning attributes and composition (Kong et al., 2016). Chen et al. (2017b) introduced the View Finding Network (VFN) which compares pairs of views to estimate the most aesthetically pleasing one. It is composed of a CNN with a ranking layer taking two views as input and predicting the more visually pleasing. The method has been trained on images gathered on Flickr and evaluated on ICDB and FCDB.

Two-step. Zhang et al. (2013, 2012) presented a cropping method in which a Region Adjacency Graph (RAG) is obtained by segmenting the entire image into small regions. Next, a region adjacency graph (graphlets) capturing training photo aesthetics has been constructed. Finally, a candidate search procedure based on probabilistic graphical models is performed and the inference of the cropping parameter is made through Gibbs sampling.

Iterative. The Aesthetic Aware Reinforcement Learning (A2-RL) framework avoids evaluating a large amount of proposals (typical of sliding window approaches) and reduces the search iterations for the best crop to a few dozen. It uses actor–critic (A3C) based reinforcement learning method to search the best cropping windows sequentially with only several candidate windows (Li et al., 2018). A2-RL includes an aesthetics aware reward and LSTM-based state representation which includes both the current and historical experience. The agent can choose among a set of 14 pre-defined actions, which can be divided into four groups: scaling actions, position translation actions, aspect ratio translation actions and termination action. The model is trained on a subset of ~9000 images from the AVA dataset belonging to one of three aesthetic quality levels (*i.e.* low, middle, or high quality). Extensive experiments have been conducted on three of the most used datasets in the state-of-the-art.

2.2.3. Hybrid methods

Another family of cropping methods is based on a two-step strategy through determining-adjusting. This family of methods avoids greedily searching against all possible sub-windows of an image as the sliding window-based methods do.

in Liu et al. (2010), the authors adopted a compound crop-and-retarget operator, which selects a subset of the image objects, whose relative positions are then adjusted by the retargeting operator. Image objects are first detected by using a saliency algorithm. Given the image objects, image prominent lines, the computed saliency map, and three aesthetics metrics (i.e. RoT, diagonal dominance, visual balance), a score evaluating the composition quality of several image regions is then predicted. Wang et al. implemented such strategy by designing a two-branch CNN for Attention Box Prediction and Aesthetics Assessment (ABP-AA) (Wang & Shen, 2017; Wang et al., 2018). First a bounding box covering the most visually important area is estimated, and then the best cropping with highest aesthetic quality is selected. The cropping algorithm is split into two cascaded stages, namely, attention-aware cropping candidates generation and aesthetic-based selection. It has been trained in parallel on two databases (i.e. SALICON (Jiang et al., 2015) and AVA (Murray et al., 2012)), then it works in a cascaded way in inference.

Kao et al. (2017) presented a two-step cropping method. In the first step, the aesthetic preservation module samples crop candidates with aesthetic score higher than a threshold from the input image. The candidate crops are then ranked by the composition module consisting of a Support Vector Machine (SVM) trained for discriminating well-composed (AVA images) and ill-composed images (random crops of the well-composed images). The cropping method returns the first 5 crops as output. The CNN-based Cascaded Cropping Regression (CCR) method (Guo et al., 2018) consists in a two-step learning approach. In the first phase, a CNN is trained for binary aesthetic quality classification using the combined AVA and CUHKPQ dataset. Secondly, the CNN features extracted from the pre-trained model are used as input for a cascaded cropping regression method, which is able to fit the image cropping information annotated by professional photographers by combining a set of weak random-ferns regressors (Dollár et al., 2010). Experimental results are reported on the ICDB. Recently, in Lu et al. (2019b) it has proposed a CNN-based method to learn the relationship between interested objects along with the corresponding visual saliency and high aesthetic quality image. An initial rectangle is first estimated by using a U-Net trained to extract visual saliency, Then, a regression neural network adapts the rectangle in order to improve

its aesthetic quality. An end-to-end automatic image cropping system learning the relationship between the interest objects and the areas with high aesthetic scores through a DNN has been presented in Lu et al. (2020). A saliency map is first predicted by a U-Net, which is then passed to a soft binarization layer to separate objects from the background. The proposed Interest Object Region (IOR) layer and the ROI warping pooling layer extract the interest object, which are finally evaluated to predict the optimal cropping region.

2.2.4. Supervised BBox estimation: Two-step

In image cropping it is important to take into account not only what remains in the cropped image, but also what is removed or modified from the original image. To this end, in Yan et al. (2013) several features have been considered to model the content and composition changes due to the cropping operation, such as foreground map estimation, color distance, texture, and sharpness. The best crop is the one with the best composition among 500 candidates obtained considering only the exclusion characteristics. The method has been trained and evaluated on the image cropping data of the proposed ICDB dataset. A meta-learning based method for the aspect-ratio-specified image cropping is presented in Li et al. (2020a). The goal of the learning process of this method is to learn cropping models for different aspect ratio requirements. In the base model, there are two parameters depending on the aspect ratio, which are determined by the meta-learners: the Aspect Ratio Specified Feature Transformation Matrix (ARS-FTM), and the Aspect Ratio Specified Pixel-Wise Predictor (ARS-PWP). When both ARS-FTM and ARS-PWP are estimated, the newly generated model can predict the cropping window of the specified aspect ratio from the image.

2.2.5. Supervised ranking: Predefined anchor boxes

The previous methods aim to find the best sub-window of the image, however the search space is very huge, for this reason alternative methods have been proposed aimed at ordering a set of predefined anchor boxes. The latter pooled from multiple image re-composition and cropping algorithms.

in Chen et al. (2017a), the authors demonstrated the effectiveness of the pairwise leaning-to-rank strategy compared to traditional image cropping techniques on the proposed FCDB dataset. Wei et al. (2018) exploited the teacher–student learning paradigm to train the View Evaluation Network (VEN) and the View Proposal Net (VPN), respectively. The VEN is a Siamese architecture trained on pair of images by using the RankingLoss as criterion. The VPN is in charge of ranking 895 predefined anchor boxes and is optimized by using the proposed pairwise ranking orders loss on all the anchor boxes. ASM-Net (Tu et al., 2020) estimated a composition-aware and saliency-aware aesthetic score map of the same size as the input image. For each crop of a set of predefined anchor boxes, the crop-level score was estimated by pooling the pixels of the map belonging to the crop. The map is obtained using a multi-scale feature extractor trained with two pairwise ranking losses to estimate which crop has the best composition and the most salient between pair of crops. in Lu et al. (2019a), the learning of photo composition has been formulated as a listwise ranking problem. The Listwise View Ranking Network (LVRN), given an image, extracts its features using a VGG16 as backbone, the refined view sampling module cuts the features related to a series of candidate views from the entire feature map, which are finally sorted based on their composition.

2.2.6. Supervised regression: Predefined anchor boxes

Still with the aim of reducing the number of candidate crops, a grid anchor based formulation of image cropping in Zeng et al. (2019), Zeng et al. (2020) has been proposed. Unlike the methods of the previous family, a very light deep cropping model trained to estimate a quality score for each candidate has been designed. The experiments conducted on their GAICD dataset have shown the effectiveness of the model based on three new types of metrics defined to reliably and comprehensively evaluate the cropping performance. In Li et al. (2020b), Li et al. improved the previous method by introducing in it a gated-based module to model mutual relations between different candidates crops.

Table 1
Backbone architecture (Sandler et al., 2018). Each line describes a sequence of 1 or more identical layers, repeated n times. All layers in the same sequence have the same number c of output channels, instead c changes from a sequence to another. The first layer of each sequence has a stride s and all others use stride 1. t represents the expansion factor. The output of the sequence in blue is used as feature volume for the first scale, the output of the sequence in red is used as feature volume for the second scale, finally the output of the gray sequence is the third feature volume.

Input	Operator	t	c	n	s
$224 \times 224 \times 3$	conv2d	–	32	1	2
$112 \times 112 \times 32$	bottleneck	1	16	1	1
$112 \times 112 \times 16$	bottleneck	6	24	2	2
$56 \times 56 \times 24$	bottleneck	6	32	3	2
$28 \times 28 \times 32$	bottleneck	6	64	4	2
$28 \times 28 \times 64$	bottleneck	6	96	3	1
$14 \times 14 \times 96$	bottleneck	6	160	3	2
$7 \times 7 \times 160$	bottleneck	6	320	1	1
$7 \times 7 \times 320$	conv2d 1×1	–	1280	1	1
$7 \times 7 \times 1280$	avgpool 7×7	–	–	1	–
$1 \times 1 \times 1280$	conv2d 1×1	–	n_classes	–	–

3. Method

The autocropping method we propose here is based on the idea that a good cropping strategy should identify a sub-region of the original image which is the best in terms of aesthetics, geometric composition, and more important, that preserves semantics of the main content of the original image. Fig. 4 shows the proposed method. The method takes inspiration from the anchor-based approaches but extends the underlying idea by considering multiple cropping criteria instead on a single one. Given an input image and a corresponding list of predefined anchor boxes, three different strategies based on aesthetics, composition and semantics are used to generate three ranked lists of the input anchor boxes. Each strategy is based on a Deep Convolutional Neural Network (CNN) specially trained to perform the assigned task. The output of the method is a final ranked list of anchor boxes that is the average of the three lists generated in the previous steps. The best crop is the head of the output ranked list which is the one with the highest score. The evaluation procedure of the different pre-defined anchor boxes is inspired by Zeng et al. (2020), while the criteria with which these are evaluated extends (Fang et al., 2014).

3.1. The cropping model

All the three strategies employed are based on the same pipeline as shown in Fig. 5 which involves: (i) a lightweight and efficient backbone based on the MobileNet-v2 architecture (Sandler et al., 2018) which takes an input image of size $N \times M$ and extracts a feature volume having spatial resolution equal to $H \times W$ and C channels using a multi-scale approach; (ii) a cropping module, based on three convolution layers, that takes the flattened version of the feature volume corresponding to each pre-defined anchor box as input and it outputs a score value.

Multi-scale feature extraction. We extract multi-scale features from the CNN in order to make features invariant with respect to the scale of objects into photographs. Specifically, we take the features to three different scales by gathering activations of MobileNet-v2 as specified in Table 1. These $H \times W \times C$ feature volumes have dimensions corresponding to $32 \times 32 \times 32$, $16 \times 16 \times 96$, and $8 \times 8 \times 320$, respectively. Since they have a different spatial resolution, we use bilinear downsampling

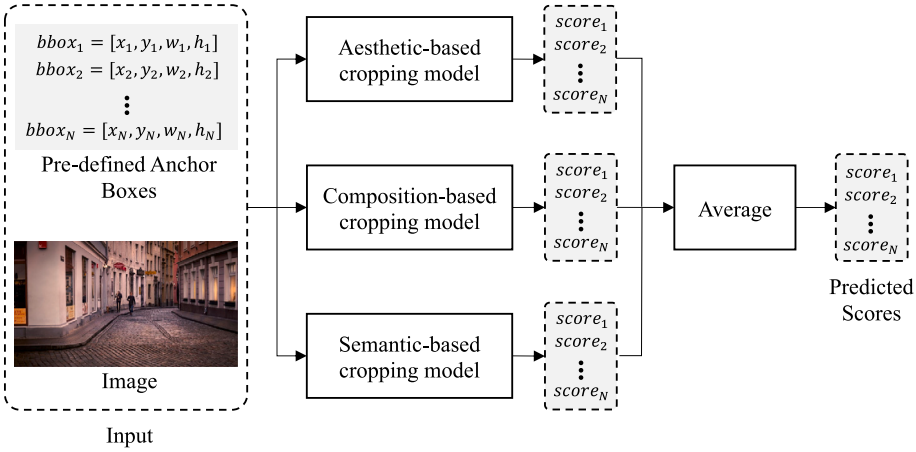

Fig. 4. Overview of the proposed cropping method. Given an image and a set of pre-defined anchor boxes as input, the method estimates a score for each box of the set by averaging the predictions of three models which models aesthetics, geometric composition, and semantics of the crops. The final scores are then ranked to select the best crop.

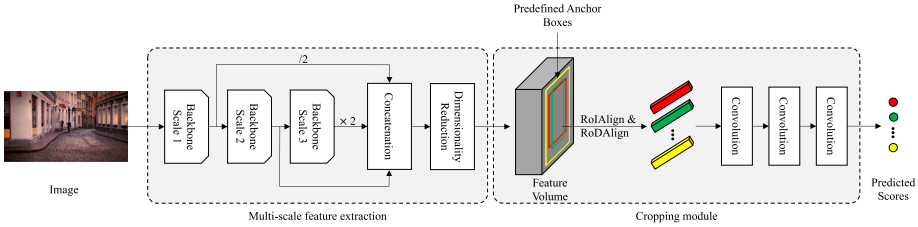

Fig. 5. Cropping model definition. For a test image the method extract multi-scale features that are scaled to the same resolution, concatenated along the channel dimension, and finally reduced on the number of channels. Given the previous feature volume and a set of pre-defined anchor boxes, the crop module employs the RoIAlign (He et al., 2017) and the RoDAlign (Zeng et al., 2019) modules to extract the regions of the volume corresponding to each box. Finally each of these feature vectors is passed in a series of convolutional blocks to finally produce a score for each box.

and upsampling to make them having the same spatial resolution. Then the three feature volumes are concatenated along the channel dimension obtaining a feature volume of shape $16 \times 16 \times 448$ (i.e. $32 + 96 + 320$ channels).

The number of channels of the feature map is then reduced to only 8 by using a 1×1 convolution for making our image cropping model very efficient and lightweight without loss of performance.

Cropping module. Given a pre-defined anchor box, this module selects two different regions from the feature volume calculated in the previous step by following the strategy proposed by Zeng et al. (2019): one contains the cropped region of the image (Region of Interest, RoI), the other contains the region to be discarded (called also the Region of Discard (RoD)). The RoIAlign operation (He et al., 2017) is used for extracting the portion of the feature map corresponding to the RoI and then the bilinear interpolation is used to resize it at a fixed spatial resolution of 9×9. The RoDAlign (Zeng et al., 2019) is instead applied to gather the Region of Discard (RoD) by zeroing the region of the whole feature map corresponding to the RoI and by using bilinear interpolation to scale the RoD feature map at the same spatial resolution of the RoI. The two feature maps are then chained along the channel dimension, flattened and fed to a stack of three fully connected layers for predicting the score of each anchor box.

Loss. The multi-scale feature extraction module is specifically pre-trained to accomplish each given task: aesthetics, composition, semantics. The cropping module is trained on the autocropping benchmark dataset used in the evaluation. To this end, a SmoothL1 loss is employed:

$$\text{SmoothL1}(g_i, p_i) = \begin{cases} 0.5(g_i - p_i)^2, & \text{if } |g_i - p_i| < 1 \\ |g_i - p_i| - 0.5, & \text{otherwise.} \end{cases} \quad (1)$$

where, g_i and p_i denote the ground-truth MOS and predicted scores of the ith candidate respectively. This loss is widely used for regression tasks because of its robustness to outliers.

3.2. Cropping models ensemble

Many deep learning-based cropping methods, consider only a single criterion to define the best crop or crops in an image. However, as we argued, a single criterion does not capture the complexity of human preferences that may be influenced by prior knowledge and backgrounds. Here, we exploit three quality criteria for a crop: image aesthetic, composition and semantic. The first criterion is widely used in the context of image cropping as its inclusion is to have a crop image that is visually pleasing. The second criterion is related to photographic aspects of the image: the crop region should have a good balance in how the elements are arranged in the image. The last criterion refers to the proper content of the image: the cropped region should preserve the same semantic information as the original image. These criteria do not cover all the possible aspects of a good image crop. Nonetheless, we considered them as to be sufficiently diverse and complementary to be used as a case study in our approach.

We model the aforementioned criteria, by pre-training the backbone architecture of the multi-scale feature extraction module for encoding image aesthetics, photographic composition, and image semantic, respectively.

The image aesthetics model is based on the formation of the MobileNet-v2 following the approach proposed in Talebi and Milanfar (2018) on the AVA (Murray et al., 2012) dataset. It consists in interpreting the distribution of human ratings of a given image as a probability distribution and in minimizing the error between this distribution and that predicted by the model by exploiting the loss of

the earth's mobile distance (EMD) loss. We train the model for 100 epochs using Stochastic Gradient Descent (SGD).

The model able to predict the basic photographic composition guidelines is obtained by training the MobileNet-v2 architecture on the KU-PCP dataset, which consists of 4244 outdoor photographs (3169 for training and 1075 for testing) (Lee et al., 2018). It has been annotated by 18 human subjects to categorize images into nine not mutually exclusive geometric composition classes: Rule of Thirds (RoT), vertical, horizontal, diagonal, curved, triangle, center, symmetric, and pattern. We train our model by using the binary cross-entropy loss and the SGD optimizer for 90 epochs with learning rate initially set to 0.001 and dropped by half every 30 epochs.

Finally, the model characterizing semantics is built by training the CNN for image categorization on the 1000 classes of the ImageNet dataset. SGD and cross-entropy loss are used for optimizing the model for 30 epochs with an initial learning rate equal to 0.1 which decays by a factor of 0.1 every 30 epochs. We randomly crop and horizontally flip training images for data augmentation.

Each of the above pre-trained backbones is then used to train three cropping models as described in Section 3.1: the aesthetics-based cropping model, the composition-based cropping model, and the semantic-based cropping model. Each of the previous models, both the backbone and the cropping module, have independent weights that are not shared.

Given an image, each cropping model predicts its own list of composition scores for a list of N pre-defined anchor boxes, namely $\mathbf{s^A} = [s_1^A, s_2^A, \dots, s_N^A]$, $\mathbf{s^C} = [s_1^C, s_2^C, \dots, s_N^C]$, and $\mathbf{s^S} = [s_1^S, s_2^S, \dots, s_N^S]$. The overall score of a box is finally obtained by calculating the average of the scores obtained by the three cropping models for that box: $s_i^{ASM} = \frac{s_i^A + s_i^C + s_i^S}{3}$. The final scores are then ranked to select the best crop for the image.

4. Experimental setup

We compare our method with the state of the art by objectively evaluating performance on GAICD and legacy datasets.

4.1. Datasets

We train and evaluate the proposed cropping method on the Grid Anchor Based Image Cropping dataset (GAICD) (Zeng et al., 2019). It has 106,860 candidate crops taken from 1236 total images. A Mean Opinion Score (MOS) that represents the composition quality is associated to each candidate crop. The images are divided into 1036 training images and 200 test images.

We also perform a comparison with the state of the art on legacy datasets well known in the community: ICDB, FLMS, FCDB. We believe that this comparison has many critical issues, which have also been pointed out in our previous work (Celona et al., 2019), but we believe it is important to include it to complete the study.

The ICDB database is a collection of 950 images gathered from the CUHKPQ dataset (Yan et al., 2013). It contains seven classes of images, i.e. animal, architecture, human, landscape, night, plant, and static. A cropped region is respectively annotated for each image by three different professional photographers. The images are taken from an existing image quality assessment dataset, the CUHKPQ dataset (Tang et al., 2013). The images are of varying aesthetic quality and are of different image categories.

The FLMS dataset consists of 500 images crawled from Flickr (Fang et al., 2014). These images have been selected for their imperfect composition and have different contents. Each image is cropped by 10 expert users on AMT who passed a strict qualification test. There is no ranking of the views. Each view is considered separately. No further details are provided in Fang et al. (2014) about this dataset.

The Flickr Cropping DataBase (FCDB) contains 1743 non-iconic images gathered from Flickr (Chen et al., 2017a). The cropping annotation for each image derives from the choices of four AMT workers who evaluated several candidate views manually drawn. 348 out of the 1743 images are adopted as test set.

Table 2
Characteristics of the image cropping datasets used in the experiments.

Dataset	Images	Views	Source	Crops	Evaluation	Ranking
GAICD	1,236	106,860	Flickr	Grid	19 experts	Yes
ICDB	950	3	CUHKPQ	Human	3 experts	No
FCDB	348	1	Flickr	Human	AMT workers	No
FLMS	500	10	Flickr	Human	10 experts	No

Table 2 summarizes the characteristics of the databases used in the experiments in terms of the number of images, the number of views i.e. the number of crops available for each image, the source from which the images were taken, how the crops have been obtained, who annotate the crops, and whether the different views are ranked by preference.

4.2. Evaluation metrics

Ranking correlation metrics. Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank-Order Correlation Coefficient (SROCC) are used for estimating prediction consistency with ground-truth MOS. First PLCC and SROCC are measured between the MOS vector of all the crops of an image and the scores of these crops estimated by the model, then the average PLCC and SROCC over the testing images are computed as final results.

Best return metrics. The practical purpose of a cropping algorithm is to return the best crops rather than accurately rank all the candidate crops, so a set of metrics have been defined to assess the ability of the models to return the best crops (Zeng et al., 2019). This new set of metrics is called the "return K of top N" accuracy. It checks on average how many of the returned K crops fall into the top-N best crops of an image denoted as $S_i(N)$. It is defined as:

$$Acc_{K/N} = \frac{1}{TK} \sum_{i=1}^{T} \sum_{k=1}^{K} True(c_{ik} \in S_i(N)), \tag{2}$$

where $True(*) = 1$ if $*$ is true, otherwise $True(*) = 0$. Following (Zeng et al., 2019) the number of returned crops K is equal to 4, and the number of best crops N is set to 5 or 10. A total of 8 accuracy indexes $Acc_{K/N}$ is obtained based on the combination of K and N.

Rank weighted best return metrics. Given that the previous metric does not distinguish the rank among the return top-N crops, a variant of the $Acc_{K/N}$ metric has been introduced. The "rank weighted return K of top-N" accuracy, which is defined as

$$Acc^{w}_{K/N} = \frac{1}{TK} \sum_{i=1}^{T} \sum_{k=1}^{K} True(c_{ik} \in S_i(N)) * w_{ij}, \tag{3}$$

where

$$w_{ij} = e^{\frac{-\beta(r_{ij}-j)}{N}}, \tag{4}$$

in which $\beta > 0$ is a scaling parameter set to 1 as in Zeng et al. (2019). The definition of the weight w_{ij} is aimed at rewarding the correct rank and being equal to 1 when the sorted rank r_{ij} matches the order of c_{ij} among the K returns.

Intersection-over-union (IoU). The intersection-over-union (IoU), also referred to as the Jaccard index, is essentially a method to quantify the percent overlap between the ground-truth candidate view and the predicted crop. Given the area of the ground-truth candidate view W_{GT} and the area of the predicted crop W, the IoU is defined as follows:

$$IoU = (W_{\mathrm{GT}} \cap W)/(W_{\mathrm{GT}} \cup W) \tag{5}$$

Boundary displacement error (BDE). The boundary displacement error computes the distance between the four edges of the ground-truth candidate view and the corresponding edges of the predicted crop. By denoting the four edges of the ground-truth candidate view and of the predicted view respectively as $B_{GT}(l)$, $B_{GT}(r)$, $B_{GT}(t)$, $B_{GT}(b)$, and $B(l)$, $B(r)$, $B(t)$, $B(b)$. The BDE is estimated as follows:

$$BDE = \sum_{j=\{l,r,u,b\}} |B_{GT}(j) - B(j)|/4, \tag{6}$$

4.3. Implementation details

We implement our method using the PyTorch framework. In the training phase, our method takes as input an image and 64 randomly selected candidate crops from the set of annotated candidates. To improve the generalization ability of the proposed method, we apply data augmentation techniques by randomly adjusting brightness, contrast, saturation, hue, and by horizontally flipping the input image and the corresponding candidate crops. Data augmentation does not affect the image composition. During the testing phase, the trained method evaluates all the annotated candidate crops and estimates a score for each one of them. During both training and test, the short side of input images is resized to 256 pixels, and pixel values are scaled to the range of [0,1] and normalized using the mean and standard deviation calculated on ImageNet. The MOS values are also normalized by removing the mean and dividing by the standard deviation across the training set. The Adam optimizer (Kingma & Ba, 2014) is used to train our models for 40 epochs with a fixed learning rate of $1e^{-4}$.

5. Results

We compare our method with previous methods which released the source code or executable program: View Finding Network (VFN) (Zeng et al., 2019), View Evaluation Network (VEN) (Wei et al., 2018), A2-RL (Li et al., 2018), and two version of GAIC: the conference version (Wei et al., 2018) and the journal version trained on the conference version of GAICD (Zeng et al., 2020).

We evaluate four variants of the proposed method, which we call Our_A, Our_C, Our_S, and Our_ACS. The first three consist of a single cropping model based on a pre-trained network on aesthetics, composition or semantics respectively. The last one represents our full method that considers all the three criteria.

5.1. Results on the GAICD

Table 3 reports the performance of the proposed method and its variants compared with the state of the art on the GAICD dataset. All the considered methods, apart from A2-RL, output scores for all the candidate crops provided by the dataset, thus for those methods we can estimate all the defined evaluation metrics. A2-RL outputs a single crop so we can only compute $\text{Acc}_{1/5}$, $\text{Acc}^{\text{w}}_{1/5}$, $\text{Acc}_{1/10}$, and $\text{Acc}^{\text{w}}_{1/10}$ for it.

From the analysis of the results reported in the Table 3, it is possible to make various considerations. First of all, our full cropping method (Our_ACS) outperform previous ones with respect to all the considered metrics.

This confirms our initial assumption that a cropping method should take into account several criteria in order to select the best crop. Having different criteria, each of which, focus on different aspects of image content and perception, is more effective than using only one or two image characteristics as in previous methods in the state-of-the-art.

This is also supported by the performance obtained by our methods leveraging either the aesthetic, the composition or the semantic. Each of these methods achieves good results but not as good as the full model. This indicates that several criteria intervene and intermix in deciding what a good crop is. The full model better fit the user preferences expressed by the MOS of the GAICD dataset.

Fig. 6 shows the qualitative comparison between our method and the state of the art. The first row of the figure contains the source images, the last line reports the ground-truth crop for each image. Each of the remaining rows exhibits the first crop returned by each method for a given image. As can be seen, our method produces crops very similar to the ground-truth. Our_ACS and GAIC crop almost in the same way but GAIC tends to preserve more information. A2-RL sometimes returns the source image without removing any distracting element. Both VFN and VEN cut important content in photos with the main subject e.g. in the photo of the cat, while outputs 16:9 crops for panoramic photos.

5.2. Computation time

In the last column of Table 3, we compare the computation time in terms of frame-per-second (FPS) on both GPU and CPU for all the considered methods. All methods are tested on the same desktop computer with an Intel Core i7-8700 CPU@3.20 GHz, 16 GB DDR4 RAM 2400 MHz, and NVIDIA GTX 1080Ti with 3580 CUDA cores. Our method, GAIC (Zeng et al., 2020), GAIC (Conf.) (Zeng et al., 2019), and VEN (Wei et al., 2018) are implemented using the PyTorch framework. A2-RL (Li et al., 2018) and VFN (Chen et al., 2017b) are implemented in Tensorflow. As can be seen, the fastest method is GAIC based on the MobileNet-v2 at 200 FPS on GPU and 6 FPS on CPU. Our method is the second fastest among competitors at 183 FPS on GPU and 2 FPS on CPU. The other methods are much slower because they employ heavy CNN architectures (GAIC (Conf.)), individually process each crop (VFN and VEN), or repeatedly update the cropping window (A2-RL).

Previous performance for our method is estimated by running the three cropping models that make up our method in parallel. Since the architecture of the cropping models is lightweight (it includes a MobileNet-v2), it is possible to load the three models into memory and perform the forward steps in parallel, eventually averaging the estimated scores from the three models. Running the three cropping models sequentially decreases the performance of our method, namely 58 FPS on GPU and 0.614 FPS on CPU.

5.3. Ablation study

In this ablation study, we firstly evaluate different ways to combine scores predicted by aesthetics **A**, composition **C** and semantic-based **S** models. We then compare the joint use of aesthetics, composition and semantic **ACS** with variants such as **AC**, **CS**, etc. All the evaluation are carried out on the GAICD dataset.

Combining cropping model predictions. As described in Section 3.2, given an image, each cropping model predicts its own list of composition scores for a list of N pre-defined anchor boxes, namely $\mathbf{s^A} = [s^A_1, s^A_2, \dots, s^A_N]$, $\mathbf{s^C} = [s^C_1, s^C_2, \dots, s^C_N]$, and $\mathbf{s^S} = [s^S_1, s^S_2, \dots, s^S_N]$. The overall score of a box can be obtained by calculating the minimum, the maximum, the average, or the weighted average of the scores obtained by the three cropping models. The four combining approaches considered can be formally described as:

$$s_i^{min} = \min(s_i^A, s_i^C, s_i^S), \tag{7}$$

$$s_i^{max} = \max(s_i^A, s_i^C, s_i^S), \tag{8}$$

$$s_i^{avg} = \frac{s_i^A + s_i^C + s_i^S}{3}, \tag{9}$$

$$s_i^{wavg} = \frac{w^A s_i^A + w^C s_i^C + w^S s_i^S}{3}. \tag{10}$$

Here, w^A, w^C, and w^S are three weights balancing the contribution of the score predicted by each cropping model. The previous weights are optimized on the training set of the dataset using the least square method. The best results for all the considered metrics are obtained by averaging cropping model predictions. In particular, its $\text{Acc}_{1/5}$ equal

Table 3
Comparison with state-of-the-art methods on GAICD (Zeng et al., 2019). The "–" means that the result is not available.

Method	Backbone	PLCC	SROCC	$Acc_{1/5}$	$Acc_{4/5}$	$Acc^{w}_{1/5}$	$Acc^{w}_{4/5}$	$Acc_{1/10}$	$Acc_{4/10}$	$Acc^{w}_{1/10}$	$Acc^{w}_{4/10}$	FPS (GPU)	FPS (CPU)
A2-RL (Li et al., 2018)	AlexNet	–	–	24.5	–	15.6	–	41.0	–	26.9	–	4	0.047
VFN (Chen et al., 2017b)	AlexNet	0.470	0.450	27.0	24.6	16.8	11.1	39.0	37.3	25.9	19.1	0.4	0.004
VEN (Wei et al., 2018)	VGG16	0.653	0.621	40.5	36.8	20.0	12.8	54.0	48.4	30.0	23.8	0.3	0.002
GAIC (Conf.) (Zeng et al., 2019)	VGG16	0.762	0.735	53.5	46.6	37.6	30.0	71.5	65.5	53.7	46.9	136	1.278
GAIC (Zeng et al., 2020)	VGG16	0.782	0.758	59.0	48.6	37.2	34.2	74.0	67.4	53.7	50.7	123	1.189
GAIC (Zeng et al., 2020)	MobileNetV2	0.806	0.783	62.5	52.5	39.6	36.2	78.5	72.3	56.9	54.4	**200**	**6**
Our_A	MobileNetV2	0.797	0.773	60.5	50.2	39.6	35.2	78.5	70.5	56.3	52.4	**200**	**6**
Our_C	MobileNetV2	0.807	0.781	62.5	52.0	39.6	37.0	78.5	70.8	56.5	53.6	**200**	**6**
Our_S	MobileNetV2	0.789	0.764	62.0	50.5	41.4	36.0	79.0	69.3	57.3	52.2	**200**	**6**
Our_{ACS}	MobileNetV2	**0.822**	**0.798**	**65.5**	**54.4**	**44.9**	**39.2**	**83.5**	**72.5**	**61.5**	**55.5**	183	2

Fig. 6. Qualitative comparison of returned top-1 crops by different methods on GAICD test images.

Table 4
Results given by the combination of the three considered criteria, namely: A - aesthetics, C - composition, and S - semantics. In each column, the best and second-best results are marked in **boldface** and underlined, respectively.

Method	PLCC	SROCC	$Acc_{1/5}$	$Acc_{4/5}$	$Acc^{w}_{1/5}$	$Acc^{w}_{4/5}$	$Acc_{1/10}$	$Acc_{4/10}$	$Acc^{w}_{1/10}$	$Acc^{w}_{4/10}$
Our_{A}	0.797	0.773	60.5	50.2	39.6	35.2	78.5	70.5	56.3	52.4
Our_{C}	0.807	0.781	62.5	52.0	39.6	37.0	78.5	70.8	56.5	53.6
Our_{S}	0.789	0.764	62.0	50.5	41.4	36.0	79.0	69.3	57.3	52.2
Our_{AC}	0.818	0.794	63.5	54.2	42.2	**39.2**	79.5	72.2	59.0	55.2
Our_{CS}	0.817	0.793	64.0	53.9	44.3	38.3	82.5	72.1	60.5	54.8
Our_{AS}	0.809	0.785	**65.5**	52.4	**45.2**	37.4	82.0	71.4	60.0	54.3
Our_{ACS}	**0.822**	**0.798**	**65.5**	**54.4**	44.9	**39.2**	**83.5**	**72.5**	**61.5**	**55.5**

Fig. 7. Qualitative comparison returned top-1 crops obtained by our crop models trained on the three criteria.

to 65.5% is 2% higher than the second approach consisting in taking the maximum among the three scores per candidate crop, and 6% with respect to worst performance achieved by taking the minimum. On average, the second best approach is to take the maximum of the predicted scores, while the weighted average performs worse than the average and the maximum probably due to overfitting on the training set.

Criteria for evaluating candidate crop regions. To demonstrate that the set of the three considered criteria (that is, aesthetics, composition, and semantics) allows selecting the best crop, we alternate the criterion or the criteria for evaluating the crops. More in detail, after measuring the performance obtained by considering one criterion at a time for the rank of candidate crops, we also analyze pairs of criteria, namely aesthetics+composition (named Our_{AC}), composition+semantics (Our_{CS}),

Table 5
Comparison with state-of-the-art methods on the ICDB (Yan et al., 2013) dataset in terms of IoU and BDE.

Method	Phot. 1		Phot. 2		Phot. 3	
	IoU ↑	BDE ↓	IoU ↑	BDE ↓	IoU ↑	BDE ↓
Baseline	0.823	0.046	0.830	0.046	0.808	0.050
VFN (Chen et al., 2017b)	0.764	0.064	0.753	0.068	0.733	0.074
VEN (Wei et al., 2018)	0.781	–	0.770	–	0.753	–
GAIC (Zeng et al., 2020)	0.785	0.055	0.760	0.063	0.763	0.061
A2-RL (Li et al., 2018)	0.802	0.052	0.796	0.053	0.790	0.053
ABP-AA (Wang & Shen, 2017)	0.813	0.030	0.806	0.032	0.816	0.032
Our_{ACS}	0.784	0.056	0.759	0.063	0.763	0.061

and aesthetics+semantics (Our_{AS}). The overall score of a box is obtained by calculating the average of the scores obtained by the two cropping models.

Table 4 reports the results for the previous experiments on the test set of GAICD dataset. We would like to highlight that the combination of the three criteria, Our_{ACS}, obtains the best performance on all metrics apart from $\text{Acc}^{\text{w}}_{1/5}$. The second best results are achieved by the method considering both aesthetics and composition, Our_{AC}. Finally, the worst results are obtained by considering only one criterion, in particular the one based on semantics (Our_{S}). Looking at the scores obtained by this latest cropping model for each image, it is possible to see that they are not very different from each other. This is probably due to the fact that the semantic features extracted for the candidate crops are unable to discriminate high from low quality crops.

Fig. 7 shows the qualitative comparison of the crops obtained using the cropping models based on a single criterion, namely aesthetics (Our_{A}), composition (Our_{C}), and semantics (Our_{S}), and the combination of the three criteria (Our_{ACS}). Some considerations can be made, on the basis of the results. First of all, for the sampled images, the crops obtained for Our_{A}, Our_{C}, and Our_{S} are different. This confirms that cropping models based on different criteria output different crop ranks. Second, the Our_{ACS} crop is sometimes present among Our_{A}, Our_{C}, and Our_{S}. For example, for the first column image, Our_{ACS} is identical to Our_{A}, for the third column image, Our_{ACS} is Our_{C}. It could happen that the combination of the three criteria can re-rank the candidate crops producing as best crop a candidate that does not excel for the criteria taken individually. This is illustrated in the case of the sample image in the last column.

5.4. Results on legacy datasets

Tables 5, 6 and 7 reports on the comparison between the proposed method and the state of the art on the ICDB, FLMS and FCDB, respectively. Following the original works, for comparison, here the metrics used are the IoU and BDE.

With respect to the ICDB dataset, Table 5 shows that the attention-based approach ABP-AA (Wang & Shen, 2017) achieves the best overall results followed by the aesthetic-based A2-RL method. The GAIC and Our methods perform similarly and are in a third position. We can also see that, depending on which photographer GT is considered, the performance varies. Specifically, Photographer 1 seems to have a GT different from Photographers 2 and 3 whose results are similar. This could suggests that experts may have quite different opinions on what to consider relevant for a crop image.

Table 6 shows the results obtained on the FLMS dataset. Similarly to the ICDB dataset, the only two measures considered are the IoU and BDE. In this case, we see that the best results are obtained by the VEN approach (Wei et al., 2018), followed by AIC (Wang et al., 2018), and then by A2-RL. Our approach has similar results to GAIC. On the overall all the methods, with the exception of Chen et al. (2016) have results above 0.81 in IoU and are not very dissimilar.

Finally, on the FCDB dataset (Table 7), the best performance are achieved by the VEN (Wei et al., 2018) method. The other methods, with the exception of RankSVM (Chen et al., 2017a), have very similar results both in terms of IoU and BDE.

Table 6
Image cropping results obtained for the FLMS dataset (Fang et al., 2014).

Method	IoU ↑	BDE ↓
Baseline	0.586	0.116
Chen et al. (2016)	0.640	0.075
ABP-AA (Wang & Shen, 2017)	0.810	0.057
GAIC (Zeng et al., 2020)	0.817	0.046
A2-RL (Li et al., 2018)	0.820	–
AIC (Wang et al., 2018)	0.830	0.052
VEN (Wei et al., 2018)	0.837	0.041
Our_{ACS}	0.818	0.045

Table 7
The IoU (and BDE) obtained for FCDB (Chen et al., 2017a).

Method	IoU ↑	BDE ↓
Baseline	0.636	0.100
RankSVM (Chen et al., 2017a)	0.602	0.106
AIC (Wang et al., 2018)	0.650	0.080
A2-RL (Li et al., 2018)	0.663	0.089
GAIC (Zeng et al., 2020)	0.665	0.085
VFN (Chen et al., 2017b)	0.684	0.084
VEN (Wei et al., 2018)	0.735	0.072
Our_{ACS}	0.682	0.083

Observing the results in the three previous tables, we can see that there is no method that consistently perform well on all the datasets. Each dataset presents a different best method and also the other approaches rank differently on each dataset. The main reason for this behavior is that, as demonstrate in our previous work (Celona et al., 2019), each dataset models the cropping problem differently. This means that the subjects who created the GT for images use different rationals to select their best crop region. Methods that consider only one dataset as their benchmark dataset may be biased towards that specific data and thus do not perform similarly well on other datasets having different rationals. The contents of the images play an important role in deciding how to crop an image. Image datasets with very similar content are not very useful for designing a general purpose cropping method. For example, in Celona et al. (2019), we have shown that on some datasets, a dummy cropping strategy (indicated as baseline in Tables 5, 6 and 7), i.e. considering the whole image as cropping region, is able to achieve comparable results to more complex methods. These show that there are biases in some of the cropping datasets usually used in evaluating cropping algorithms. For these reasons, to evaluate the effectiveness of the cropping methods, we have also considered performing subjective tests with panels of users. The tests are aimed at collecting user preferences on the cropping results across different datasets, in a comparative manner.

5.5. User preference study

We conduct two user studies to compare the output of different cropping methods. For each source image, we show the human subject the source image as reference, and several cropping results obtained by different methods displayed in random order. The human subject

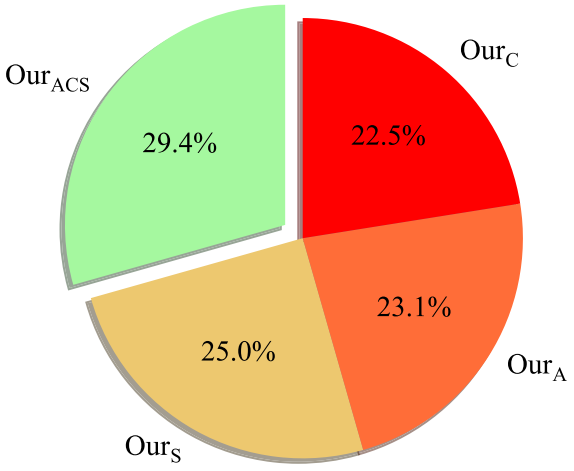

Fig. 8. User rating statistics for image versions obtained using the four variants of our cropping method.

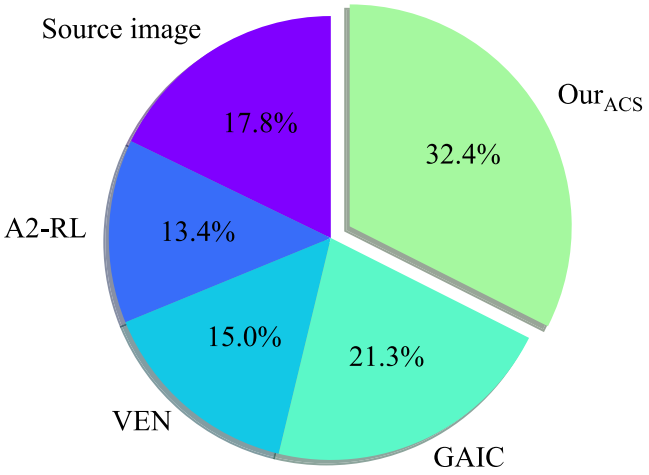

Fig. 9. User rating statistics for image versions obtained using different cropping methods.

is asked to choose the best view from those compared for each image. To make the comparison fair, we randomly selected 25 images from each of the four considered datasets (thus a total of 100 images). We invited 20 subjects to participate in the user study: 10 randomly selected participants were involved in the first subjective study, while the remaining 10 took part in the second user study.

In the first subjective study, we compare the cropping results obtained by our four variants, namely Our$_A$, Our$_C$, Our$_S$, and Our$_{ACS}$. In Fig. 8 we see that Our$_{ACS}$ is the most selected method with 29.4%, followed by Our$_S$ with 25%, Our$_A$ with 23.1%, and finally Our$_C$ with 22.5%. The results shows that the users prefer the cropping results produced by the method that considers all the three criteria together. This support our original idea. When considering a single criterion, the user's preferences are for semantics rather than aesthetics or composition. This suggests that, for users, it is important that the cropped region retain its original semantics. However, the very similar percentages obtained by Our$_S$, Our$_C$, and Our$_A$ indicate that different users have different opinions on what constitutes a good crop. This reinforces our idea of expanding the set of criteria that must be taken into account in cropping algorithms.

In the second study we compare the cropping results obtained by three state-of-the-art methods under their default settings (i.e. A2-RL (Li et al., 2018), VEN (Wei et al., 2018), GAIC (Zeng et al., 2020)), our best method from the previous experiment (Our$_{ACS}$), and the source image itself. Fig. 9 shows the statistics of the subjective study. As it can be seen, our method collected more votes than the others. Specifically, it received 32.4% votes, outperforming the second method which is GAIC by a large margin (11%). Measuring the statistics of the participants' opinions for each database separately confirms not only that our method works better than the others on GAICD, but that this also applies to the other three databases considered. This last aspect is very interesting because the VEN, which is the method to obtain the best performances on both FCDB and FLMS (see Tables 6 and 7), for our ten participants it is the third choice on these two databases after the GAIC and the proposed method.

It is interesting to note that about 18% of the times, the original, uncropped, image is chosen by the subjects. By analyzing the selections, we found out that most of these choices are made on images from the ICDB and FCDB datasets. Upon inspection of these images, it emerged that most of them depict a prominent subject encompassing the entire image. According to the users, this makes it difficult to identify a

Fig. 10. Two images of those selected for subjective studies showing prominent subjects that do not allow cropping without losing relevant regions or worsening the composition.

meaningful crop without losing relevant parts of the original image. Fig. 10 shows two examples of such images. In particular, in the second row of the figure, we show the images cropped by our method. For both images, our method selected candidate crops where edge pixels are excluded. Such results worsen the quality of the composition, especially for the image on the left, which justifies the fact that the subjects preferred the original image.

6. Conclusion

In this paper we proposed a cropping method that evaluates the goodness of a candidate cropping region leveraging different quality criteria. Our method is based on a combination of efficient and lightweight neural networks. Each network was specifically designed to evaluate the candidate crops with respect to a criterion. Specifically, for a crop region, we considered its aesthetic, overall composition, and semantics. The experimental results demonstrate that the combination of the three criteria together allows to select the most visually appealing crop region and to outperform state-of-the-art methods. A more in-depth analysis shows that, of the three cropping models, the one based on the composition obtains the best performance, while the one based solely on semantics obtains the worst results. The previous results probably indicate that between the two criteria, composition with respect to semantics has a greater ability to discriminate between high and low quality crops.

The analysis of the computational efficiency of the methods highlights how the efficient network architecture used to develop the proposed method makes it highly competitive with respect to the other methods. We also performed subjective experiments that corroborated our initial assumption that different quality criteria play an important role in determining the best crop within an image.

The experimental findings demonstrate that the combination of the three criteria is effective in learning a generic or universal cropping method. Our method can be easily extended by adding more networks to evaluate other criteria in order to further improve the performance. In future development, we therefore intend to consider other criteria, such as salience. At the same time, given that individual user's preference may differ from that of the general user, we intend to exploit and combine the criteria to better model the tastes of the individual user for the design of a personalized image cropping approach.

CRediT authorship contribution statement

Luigi Celona: Conceptualization, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Gianluigi Ciocca:** Conceptualization, Methodology, Validation, Writing – original draft, Writing – review & editing. **Paolo Napoletano:** Conceptualization, Methodology, Validation, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Celona, L., Ciocca, G., Napoletano, P., & Schettini, R. (2019). Autocropping: A closer look at benchmark datasets. In *International conference on image analysis and processing* (pp. 315–325). Springer.

Chen, J., Bai, G., Liang, S., & Li, Z. (2016). Automatic image cropping: A computational complexity study. In *Conference on computer vision and pattern recognition (CVPR)* (pp. 507–515). IEEE.

Chen, Y.-L., Huang, T.-W., Chang, K.-H., Tsai, Y.-C., Chen, H.-T., & Chen, B.-Y. (2017). Quantitative analysis of automatic image cropping algorithms: A dataset and comparative study. In *Winter conference on applications of computer vision (WACV)* (pp. 226–234). IEEE.

Chen, Y.-L., Klopp, J., Sun, M., Chien, S.-Y., & Ma, K.-L. (2017). Learning to compose with professional photographs on the web. In *International conference on multimedia* (pp. 37–45). ACM.

Chen, L.-Q., Xie, X., Fan, X., Ma, W.-Y., Zhang, H.-J., & Zhou, H.-Q. (2003). A visual attention model for adapting images on small displays. *Multimedia Systems*, *9*(4), 353–364.

Cheng, B., Ni, B., Yan, S., & Tian, Q. (2010). Learning to photograph. In *International conference on multimedia* (pp. 291–300). ACM.

Cho, D., Park, J., Oh, T.-H., Tai, Y.-W., & So Kweon, I. (2017). Weakly-and self-supervised learning for content-aware deep image retargeting. In *International conference on computer vision (ICCV)* (pp. 4558–4567). IEEE.

Ciocca, G., Cusano, C., Gasparini, F., & Schettini, R. (2007). Self-adaptive image cropping for small displays. *IEEE Transactions on Consumer Electronics*, *53*(4), 1622–1627.

Deng, Y., Loy, C. C., & Tang, X. (2018). Aesthetic-driven image enhancement by adversarial learning. In *International conference on multimedia* (pp. 870–878). ACM.

Dollár, P., Welinder, P., & Perona, P. (2010). Cascaded pose regression. In *Conference on computer vision and pattern recognition (CVPR)* (pp. 1078–1085). IEEE.

Esmaeili, S. A., Singh, B., & Davis, L. S. (2017). Fast-at: Fast automatic thumbnail generation using deep neural networks. In *Conference on computer vision and pattern recognition (CVPR)* (pp. 4622–4630). IEEE.

Fang, C., Lin, Z., Mech, R., & Shen, X. (2014). Automatic image cropping using visual composition, boundary simplicity and content preservation models. In *International conference on multimedia* (pp. 1105–1108). ACM.

Guo, G., Wang, H., Shen, C., Yan, Y., & Liao, H.-Y. M. (2018). Automatic image cropping for visual aesthetic enhancement using deep neural networks and cascaded regression. *Transactions on Multimedia*, *20*(8), 2073–2085.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In *International conference on computer vision* (pp. 2961–2969). IEEE.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259.

Jiang, M., Huang, S., Duan, J., & Zhao, Q. (2015). Salicon: Saliency in context. In *Conference on computer vision and pattern recognition (CVPR)* (pp. 1072–1080). IEEE.

Kao, Y., He, R., & Huang, K. (2017). Automatic image cropping with aesthetic map and gradient energy map. In *International conference on acoustics, speech and signal processing (ICASSP)* (pp. 1982–1986). IEEE.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kong, S., Shen, X., Lin, Z., Mech, R., & Fowlkes, C. (2016). Photo aesthetics ranking network with attributes and content adaptation. In *European conference on computer vision* (pp. 662–679). Springer.

Lee, J.-T., Kim, H.-U., Lee, C., & Kim, C.-S. (2018). Photographic composition classification and dominant geometric element detection for outdoor scenes. *Journal of Visual Communication and Image Representation*, *55*, 91–105.

Li, D., Wu, H., Zhang, J., & Huang, K. (2018). A2-RL: Aesthetics aware reinforcement learning for image cropping. In *Conference on computer vision and pattern recognition (CVPR)* (pp. 8193–8201). IEEE.

Li, Z., & Zhang, X. (2019). Collaborative deep reinforcement learning for image cropping. In *International conference on multimedia and expo (ICME)* (pp. 254–259). IEEE.

Li, D., Zhang, J., & Huang, K. (2020). Learning to learn cropping models for different aspect ratio requirements. In *Conference on computer vision and pattern recognition (CVPR)* (pp. 12685–12694). IEEE.

Li, D., Zhang, J., Huang, K., & Yang, M.-H. (2020). Composing good shots by exploiting mutual relations. In *Conference on computer vision and pattern recognition (CVPR)* (pp. 4213–4222). IEEE.

Liu, L., Chen, R., Wolf, L., & Cohen-Or, D. (2010). Optimizing photo composition. *Computer Graphics Forum*, *29*(2), 469–478.

Lu, W., Xing, X., Cai, B., & Xu, X. (2019). Listwise view ranking for image cropping. *IEEE Access*, *7*, 91904–91911.

Lu, P., Zhang, H., Peng, X., & Jin, X. (2020). Learning the relation between interested objects and aesthetic region for image cropping. *IEEE Transactions on Multimedia*, 1.

Lu, P., Zhang, H., Peng, X., & Peng, X. (2019). Aesthetic guided deep regression network for image cropping. *Signal Processing: Image Communication*, *77*, 1–10.

Murray, N., Marchesotti, L., & Perronnin, F. (2012). Ava: A large-scale database for aesthetic visual analysis. In *Conference on computer vision and pattern recognition (CVPR)* (pp. 2408–2415). IEEE.

Nishiyama, M., Okabe, T., Sato, Y., & Sato, I. (2009). Sensation-based photo cropping. In *International conference on multimedia* (pp. 669–672). ACM.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *Conference on computer vision and pattern recognition (CVPR)* (pp. 4510–4520).

Stentiford, F. (2007). Attention based auto image cropping. In *International Conference on Computer Vision Systems*.

Talebi, H., & Milanfar, P. (2018). NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, *27*(8), 3998–4011.

Tang, X., Luo, W., & Wang, X. (2013). Content-based photo quality assessment. *IEEE Transactions on Multimedia*, *15*(8), 1930–1943.

Tu, Y., Niu, L., Zhao, W., Cheng, D., & Zhang, L. (2020). Image cropping with composition and saliency aware aesthetic score map.. In *AAAI* (pp. 12104–12111).

Wang, W., & Shen, J. (2017). Deep cropping via attention box prediction and aesthetics assessment. In*Conference on computer vision and pattern recognition (CVPR)*(pp. 2186–2194).

Wang, W., Shen, J., & Ling, H. (2018). A deep network solution for attention and aesthetics aware photo cropping. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(7), 1531–1544.

Wang, W., Shen, J., Yu, Y., & Ma, K.-L. (2016). Stereoscopic thumbnail creation via efficient stereo saliency detection. *IEEE Transactions on Visualization and Computer Graphics*, *23*(8), 2014–2027.

Wei, Z., Zhang, J., Shen, X., Lin, Z., Mech, R., Hoai, M., & Samaras, D. (2018). Good view hunting: Learning photo composition from dense view pairs. In *Conference on computer vision and pattern recognition (CVPR)* (pp. 5437–5446). IEEE.

Yan, J., Lin, S., Kang, S. B., & Tang, X. (2013). Learning the change for automatic image cropping. In *Conference on computer vision and pattern recognition (CVPR)* (pp. 971–978). IEEE.

Zeng, H., Li, L., Cao, Z., & Zhang, L. (2019). Reliable and efficient image cropping: A grid anchor based approach. In *Computer vision and pattern recognition (CVPR)* (pp. 5949–5957). IEEE.

Zeng, H., Li, L., Cao, Z., & Zhang, L. (2020). Grid anchor based image cropping: A new benchmark and an efficient model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1.

Zhang, L., Song, M., Yang, Y., Zhao, Q., Zhao, C., & Sebe, N. (2013). Weakly supervised photo cropping. *IEEE Transactions on Multimedia*, *16*(1), 94–107.

Zhang, L., Song, M., Zhao, Q., Liu, X., Bu, J., & Chen, C. (2012). Probabilistic graphlet transfer for photo cropping. *IEEE Transactions on Image Processing*, *22*(2), 802–815.