Image Cropping with Spatial-aware Feature and Rank Consistency

Chao Wang, Li Niu*, Bo Zhang, Liqing Zhang*
MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University
{wangchaojfj,ustcnewly,bo-zhang}@sjtu.edu.cn, zhang-lq@cs.sjtu.edu.cn

Abstract

*Image cropping aims to find visually appealing crops in an image. Despite the great progress made by previous methods, they are weak in capturing the spatial relationship between crops and aesthetic elements (*e.g.*, salient objects, semantic edges). Besides, due to the high annotation cost of labeled data, the potential of unlabeled data awaits to be excavated. To address the first issue, we propose spatial-aware feature to encode the spatial relationship between candidate crops and aesthetic elements, by feeding the concatenation of crop mask and selectively aggregated feature maps to a light-weighted encoder. To address the second issue, we train a pair-wise ranking classifier on labeled images and transfer such knowledge to unlabeled images to enforce rank consistency. Experimental results on the benchmark datasets show that our proposed method performs favorably against state-of-the-art methods.*

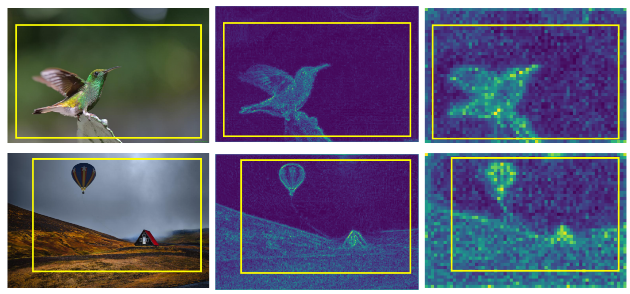

Figure 1. Two examples of the spatial relationship between crops (yellow bounding box) and aesthetic elements (*e.g.*, semantic edges and salient objects). The first column shows the source images, and the second (*resp.*, third) column shows their low-level (*resp.*, high-level) feature maps extracted by a pre-trained MobileNetv2 [39] network with channel-wise max pooling. It can be seen that low-level feature maps emphasize semantic edges and high-level feature maps highlight salient objects.

1. Introduction

The task of image cropping aims to find good crops in an image that can improve the image quality and meet aesthetic requirement. Image cropping is a prevalent and critical operation in numerous photography-related applications like image thumbnailing, view recommendation, and camera view adjustment suggestion.

Many Researchers [2, 4–7, 12, 21, 23, 36, 43, 46, 52, 54, 60, 62, 63] have studied automatic image cropping in the past decades with the goal to reduce the workload of manual cropping. Earlier works [2, 3, 12, 31, 43, 44] mainly used saliency detection [49, 59] to detect salient objects and crop around salient objects. Another group of methods [6, 12, 26, 33, 54, 62] designed hand-crafted features to represent specific composition rules in photography. With the construction of moderate-sized image cropping datasets [4, 52, 54, 56], recently proposed image cropping methods [4, 5, 7, 21, 23, 36, 52, 56, 57, 63] are usually data-driven manner and directly learn how to crop visually appealing views from the labeled data. Although these approaches have achieved impressive improvement on image cropping task, there still exist some drawbacks which will be discussed below.

One problem is that when considering the spatial relationship between crops and aesthetic elements (*e.g.*, salient objects, semantic edges), which is very critical for image cropping, previous methods usually designed some intuitive rules. For example, the crop should enclose the salient object [2, 43, 44], or should not cut through the semantic edges [2, 54]. However, these hand-crafted rules did not consider the spatial layout of all aesthetic elements as a whole, and may not generalize well to various scenes because the rules designed for specific subjects can not cover complex image cropping principles [10].

In this work, we explore learnable spatial-aware features, which encode the spatial relationship between crops and aesthetic elements. We observe that the feature map obtained using channel-wise max pooling can emphasize some aesthetic elements. In Figure 1, we show several pooled feature maps from MobileNetv2 [39], from which

*Corresponding author

it can be seen that the low-level feature maps emphasize semantic edges (*e.g.*, the outlines of semantic objects and regions) and the high-level feature maps emphasize salient objects (*e.g.*, bird, balloon). With concatenated feature maps from different layers, we learn channel attention [16] to select important layers. The weighted feature maps are concatenated with candidate crop masks and sent to a light-weighted encoder to produce spatial-aware features. The extracted spatial-aware features encode the spatial relationship between candidate crops and aesthetic elements without being limited by any hand-crafted rules.

Another problem is that the cost of crop annotation is very high and the performance is limited by the scale of the annotated training set. Therefore, some previous works explored how to utilize unlabeled data to improve the cropping performance. For example, VFN [5] collects unlabeled professional photographs from public websites and perform pairwise ranking based on the assumption that the entire image has higher aesthetic quality than any of its crops. However, such assumption does not always hold obviously. VPN [52] used a pre-trained network VEN [52] to predict aesthetic scores for the crops from unlabeled images, which function as pseudo labels to supervise training a new network. However, the predicted pseudo labels may be very noisy and provide misleading guidance.

In this work, we explore transferring ranking knowledge from labeled images to unlabeled images. Specifically, given two annotated crops from a labeled image, we learn a binary pairwise ranking classifier to judge which crop has higher aesthetic quality, by sending the concatenation of two crop features to a fully connected layer. We expect that the knowledge of comparing the aesthetic quality of two crops with similar content could be transferred to unlabeled data. Given two unannotated crops from an unlabeled image, we can obtain two types of ranks. On the one hand, we can rank them according to the predicted crop-level scores. On the other hand, we can employ the pairwise ranking classifier to get the rank. Then, we enforce two types of ranks to be consistent.

We conduct experiments on GAICD [57] and FCDB [4] dataset. For unlabeled images, we use unlabeled test images, which falls into the scope of transductive learning. Our major contributions can be summarized as:

- We design a novel spatial-aware feature to model the spatial relationship between candidate crops and aesthetic elements.
- We propose to transfer ranking knowledge from labeled images to unlabeled images, and enforce ranking consistency on unlabeled images.
- Our proposed method obtains the state-of-the-art performance on benchmark datasets.

2. Related Work

In this section, we review the existing image cropping methods and introduce the learning paradigms using unlabeled data.

2.1. Image Cropping

From the perspective of data usage, the existing image cropping methods can be roughly classified into two mainstreams: rule-based and data-driven.

Rule-based methods usually utilize attention or aesthetic features to evaluate candidate crops. Some methods [2,3,12,24,25,27,31,40,42–44,46] argued that a good crop should attract enough attention and cover the dominant subject in an image. Most of them evaluated the candidates based on the results of saliency detection [49,59]. Other approaches [6,8,12,26,29,33,50,51,54,60–62] paid more attention to the overall composition quality of crop and some of them [6,26,33,62] designed hand-crafted features or specific rules to determine which candidate has high aesthetic quality. However, the cropped views obtained based on saliency usually lack overall composition and those methods using hand-crafted features are not robust enough to predict complex image aesthetics.

With several image cropping datasets [4,12,52,54,56,57] constructed in the past decade, most recently proposed image cropping methods [4,5,7,11,13,15,18,23,28,30,46,52, 52,56–58,63] are data-driven. The main paradigm of these approaches is to generate candidates in the first stage, then score or rank them with techniques like self-supervision [5], RoIAlign [14] and RoDAlign [56,57], knowledge distillation [52], aesthetic score map prediction [46], mutual relations modeling [23], or visual elements dependencies encoding [36]. Some other methods acquired candidate crops via reinforcement learning [21,22] and set predicting [17]. Unlike these methods, our proposed method models the spatial relationship between the crops and aesthetic elements in an image, contributing to evaluating the aesthetic quality more reasonably.

2.2. Semi-supervised/Transductive Learning

Due to the high annotation cost of labeled data, how to utilize unlabeled data is an important research topic, which involves several learning paradigms. Among them, semi-supervised learning exploits unlabeled data to construct a learner whose performance is beyond those with only labeled data [47]. Many semi-supervised methods have been proposed over the past decades, which can be roughly summarized into four categories: consistency regularization, proxy-label methods, generative models, and graph-based methods [34]. Among these categories, self-training [38, 41,55] and consistency regularization [20,45,53] are commonly used. Self-training methods use the classifier trained

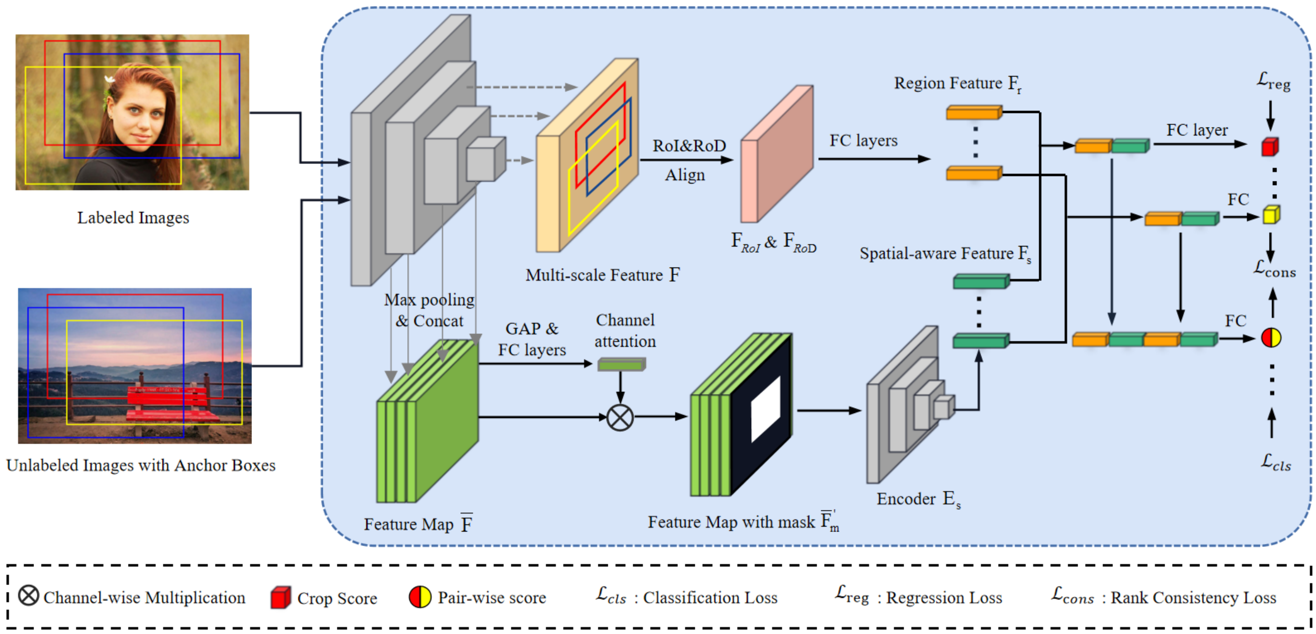

Figure 2. The flowchart of our method for image cropping with spatial-aware feature and rank consistency. The light-weighted MobileNetv2 [39] is applied as backbone to extract multi-scale features, from which region features and spatial-aware features are obtained. We train our cropping model with both labeled images and unlabeled images, during which we use annotated crop scores to supervise labeled images and rank consistency to supervise unlabeled images.

on labeled data to predict pseudo labels of unlabeled data, and then add the confident unlabeled data into training set. Consistency regularization usually enforces the prediction scores of multiple views of the same sample to be consistent. Our proposed method belongs to consistency regularization, but rank consistency between different crops in an image is specifically designed for image cropping task. Semi-supervised learning can be further divided into inductive learning and transductive learning [34, 47]. Transductive learning [1, 48] is usually applied when part of unlabeled test data are available at training time [35]. The abovementioned methods (*e.g.*, self-training, consistent regularization) can be directly applied to transductive learning.

Several existing image cropping methods [5, 52] attempted to employ unlabeled images. However, as introduced in Section 1, they either make a rigorous assumption or simply use pseudo labels. Differently, we propose a novel consistency regularization approach. In particular, we transfer aesthetic ranking knowledge from labeled data to unlabeled data and enforce rank consistency on unlabeled images.

3. Methodology

3.1. Overview

Figure 2 presents the overall flow of our proposed image cropping method with spatial-aware feature and rank consistency. Following [57], we use MobileNetv2 [39] model pre-trained on ImageNet [9] as the backbone to extract multi-scale features. We aggregate multi-scale features to obtain the region feature via RoIAlign [14] and RoDAlign [56,57], which considers not only the content in the candidate crop box but also that outside the box. Besides the region feature, we additionally extract the spatial-aware feature, which models the spatial relationship between candidate crops and aesthetic elements. We concatenate the region feature and the spatial-aware feature as crop feature followed by two branches, in which one branch predicts the aesthetic score of each candidate crop and another branch selects crop pairs for a binary pair-wise classifier to predict their relative ranks. We train our cropping model with both labeled images and unlabeled test images. For labeled images, we directly use their annotated scores for supervision. For unlabeled test images, we enforce two types of ranks to be consistent, in which one is directly from the predicted crop aesthetic scores and the other one comes from the pair-wise ranking classifier. Next, we will introduce the spatial-aware feature in Section 3.2 and rank consistency in Section 3.3.

3.2. Spatial-aware Feature

As represented in Figure 1, the low-level feature maps exhibit clear semantic edges while the high-level feature maps highlight salient objects. We exploit such observation to model the spatial relationship between candidate crops and aesthetic elements in an image, so that the model can learn the optimal placement of aesthetic elements (*e.g.*, salient objects, semantic edges) in the crop and thus locate the crop better.

To this end, we first follow [56, 57] to extract multi-scale feature maps denoted by F and obtain RoI (*resp.*, RoD) feature denoted by F_{RoI} (*resp.*, F_{RoD}) respectively with the size $h \times h$ after RoI (*resp.*, RoD) Align operations. Then we send the concatenation of them to two fully connected layers and get d_r-dim region feature F_r.

Feature Maps Activation. When extracting multi-scale feature map, we also keep different layers of feature maps. As their channel dimensions and spatial resolutions are different, we first perform max pooling along the channel dimension and then use bilinear interpolation to reshape them to the same size $H \times W$. In total, we obtain k layers of feature maps with size $H \times W \times 1$. We concatenate them along the channel dimension and denote the feature map as $\bar{F} \in \mathbb{R}^{k \times H \times W}$.

Channel Attention Block. As different layers of feature maps contain different levels of information, it is hard to decide which layers should be emphasized or suppressed. Thus, we apply the channel attention block [16] that learns the channel dependencies and performs feature recalibration automatically. As in [16], the feature map $\bar{F} \in \mathbb{R}^{k \times H \times W}$ goes through a global average pooling layer and generates channel-wise statistics, which are then delivered to two fully connected layers with activation functions to generate channel-wise weights. Finally, channel-wise weights are multiplied with feature maps $\bar{F} \in \mathbb{R}^{k \times H \times W}$, leading to $\bar{F}^{'} \in \mathbb{R}^{k \times H \times W}$. We will discuss and visualize the learned attention in Section 4.4.

Spatial Relationship Modeling. Some previous methods designed hand-crafted features to explicitly model the spatial relationship between crop and aesthetic elements (*e.g.*, exclusion features with the crop-out value and cut-through value, and compositional features considering aesthetic rules [54]). These hand-crafted features can only behave well on certain instances, so we propose to learn spatial-aware features to implicitly model the spatial relationship. Specifically, we concatenate the feature map $\bar{F}^{'} \in \mathbb{R}^{k \times H \times W}$ with one candidate crop mask (the entries within the crop bounding box are 1 and all the other entries are 0), resulting in $\bar{F}^{'}_m \in \mathbb{R}^{(k+1) \times H \times W}$. Then we send $\bar{F}^{'}_m$ to a light-weighted encoder E_s to extract the d_s-dim spatial-aware feature F_s. The Encoder E_s consists of two 5×5 convolution layers with max pooling and a fully connected layer.

Finally, we concatenate the region feature F_r and the spatial-aware feature F_s as the crop feature, and pass it to two branches for crop-level aesthetic scores prediction and pair-wise rank classification.

Optimization. We train our network in a multi-task manner. When using the labeled images, we train the aesthetic score prediction branch and the pair-wise ranking classifier at the same time, supervised by the ground-truth scores with two types of loss functions. The pair-wise ranking classifier will be discussed in Section 3.3. In the crop-level aesthetic score prediction branch, we employ smooth L_1 regression loss [37]. Given an image with N candidate crops, we denote the predicted aesthetic score and the ground-truth score of the i-th crop as $\hat{y}_i$ and y_i respectively. The regression loss is

$$\mathcal{L}_{reg} = \frac{1}{N} \sum_{i}^{N} \mathcal{L}_{s1}(y_i - \hat{y}_i), \tag{1}$$

where $\mathcal{L}_{s1}$ is the smooth L1 loss.

3.3. Rank Consistency

As image cropping aims to find good crops in the image, a lot of candidate crops need to be scored and ranked correctly. However, annotating dozens of candidate crops in an image is very expensive. Therefore, using unlabeled data is worth exploring in the image cropping task. As discussed in Section 1, previous works either make an assumption that does not always hold or use unreliable pseudo labels for knowledge distillation. In this work, we use unlabeled test images in the training stage, which falls into the scope of transductive learning [1, 48]. We explore rank consistency on unlabeled images, aiming to take advantage of the unlabeled data and use the transferred knowledge to promote cropping performance.

Pair-wise Ranking Classifier. To transfer the ranking knowledge, we first train a pair-wise ranking classifier that can distinguish the aesthetic quality of two candidate crops in the same image. As shown in Figure 2, when predicting the aesthetic score of each candidate crop, we also select crop pairs to train our pair-wise ranking classifier. Specifically, given N candidate crops with crop feature $F_{rs} = [F_r, F_s]$ in an image, we concatenate the crop features of two crops and adopt a fully connected layer to predict a score within $[0, 1]$. For a pair of two candidate crops $\{C_i, C_j\}$, the classifier output represents the probability that the aesthetic quality of C_i is better than C_j. In detail, if the value approaches 1, C_i is better than C_j, otherwise C_i is worse than C_j. For N candidate crops, we can get $(N^2 - N)/2$ crop pairs at most. However, if the margin between their scores is too small, it may confuse the model. Thus, we set score margin $\eta > 0$ to filter out the confusing pairs and get T crop pairs, from which we randomly select a fixed number of P pairs for classification. We set $\eta = 0.5$ and $P = 256$, because too many pairs increase the computational cost dramatically but bring little performance gain. The impact of hyper-parameter η and P will be discussed in Supplementary.

Optimization. We train the pair-wise ranking classifier jointly with the aesthetic score prediction branch, as discussed in Section 3.2. The loss function $\mathcal{L}_{reg}$ for the aesthetic score prediction branch has been introduced in Eqn. 1. The loss function for the pair-wise ranking classifier is the typical binary entropy loss. Specifically, we denote the classification score of the n-th crop pair $\{C_i, C_j\}$ as p_n and get their ground-truth rank label q_n according to their ground-truth scores y_i and y_j:

$$q_n = \begin{cases} 1, & if \quad y_i > y_j, \\ 0, & if \quad y_i < y_j. \end{cases} \tag{2}$$

The binary cross-entropy classification loss is

$$\mathcal{L}_{cls} = \frac{1}{P}\sum_{n=1}^{P} -q_n \cdot logp_n - (1-q_n)\cdot log(1-p_n). \tag{3}$$

When training with labeled images, the total loss is

$$\mathcal{L}_{labeled} = \mathcal{L}_{reg} + \lambda_{cls}\mathcal{L}_{cls}, \tag{4}$$

where λ_{cls} is a hyper-parameter and we set $\lambda_{cls} = 1$ via cross-validation.

Next, we transfer the ranking knowledge from labeled images to unlabeled images and impose rank consistency on unlabeled images. Given an unlabeled image, we get the pre-defined anchor boxes as in [57] and randomly select N candidate crops for training. After obtaining the crop features F_{rs}, we send them to the pair-wise ranking classifier and crop-level aesthetic score predictor. On the one hand, we use all $(N^2 - N)/2$ crop pairs for the pair-wise ranking classifier to get the rank of all crops. On the other hand, we can get another rank based on the predicted aesthetic scores. We enforce two types of ranks to be consistent using our designed consistency loss. Formally, given two crops C_i and C_j in an image, we denote their predicted aesthetic scores as $\hat{y}_i$ and $\hat{y}_j$. The output of the pair-wise ranking classifier is denoted as p_n. The consistency loss is defined as

$$\mathcal{L}_{cons} = \frac{2}{(N^2-N)}\sum_{i=1}^{N}\sum_{j=i+1}^{N} l(C_i, C_j), \tag{5}$$

where

$$l\left(C_i, C_j\right) = \max\left\{0, \delta + sign(p_n - 0.5)(\hat{y}_j - \hat{y}_i)\right\}, \tag{6}$$

in which $sign(\cdot)$ is the standard sign function and δ is a margin set as 0.1 via cross-validation. When $p_n > 0.5$ (*resp.*, $p_n < 0.5$), $\hat{y}_i$ (*resp.*, $\hat{y}_j$) is expected to exceed $\hat{y}_j$ (*resp.*, $\hat{y}_i$) by a margin δ.

So far, the total loss function can be summarized as

$$\mathcal{L}_{total} = \mathcal{L}_{labeled} + \mathcal{L}_{cons}, \tag{7}$$

in which $\mathcal{L}_{labeled}$ is trained with labeled images and $\mathcal{L}_{cons}$ is trained with unlabeled images.

4. Experiments

4.1. Datasets and Evaluation Metrics

We mainly conduct experiments on the journal version of the GAICD [57] dataset, which extended the number of source images to 3,336 (2,636 for training, 200 for validation and 500 for testing) with 288,069 labeled crops, 1,100 more images compared with its conference version [56]. We use the metrics proposed in [56], including average Spearman's rank-order correlation coefficient ($\overline{SRCC}$), average Pearson correlation coefficient ($\overline{PCC}$), and return K of top-N accuracy $ACC_{K/N}$. $\overline{PCC}$ evaluates the linear correlation between the predicted scores and the ground-truth, whereas $\overline{SRCC}$ measures the ranking order correlation which is sometimes more important in image cropping task. Following [57], we set N to 5 or 10 and K to 1, 2, 3 and 4, and get 8 return K of top-N accuracy metrics $Acc_{1/5}$, $Acc_{2/5}$, $Acc_{3/5}$, $Acc_{4/5}$, $Acc_{1/10}$, $Acc_{2/10}$, $Acc_{3/10}$, $Acc_{4/10}$. We also report their average results as $\overline{Acc_5}$, $\overline{Acc_{10}}$.

Besides the GAICD dataset, we use FCDB [4] dataset with 348 test images to evaluate our model as well. We report the Intersection-over-Union (IoU) and Boundary-displacement-error (Disp) for comparison with other approaches, even though the reliability of these two metrics is arguable [56].

4.2. Implementation Details

Following recent approaches [36, 57], we employ efficient MobileNetv2 [39] as the backbone and reduce the channel of the multi-scale feature maps to 32 with 1×1 convolution. The RoI&RoD Align resolution h is fixed to 9 as in [57]. We send all layers of the feature maps extracted from the backbone to generate spatial-aware features, that is, the number of layers k is 19. We set $H \times W$ as 64×64, and set $d_r = d_s = 256$.

During training, we resize the short side length of the source image to 256 while keeping the aspect ratio. Data augmentation like randomly horizontal flipping and photometric distorting (*e.g.*, brightness, contrast, and saturation) are employed for better generalization. We randomly select $N = 64$ candidate crops of an image as a batch for training and leverage all candidates in the test stage. We train the whole network end-to-end by using the Adam [19] optimizer with a weight decay of $1e^{-4}$ for 60 epochs. The learning rate is set to $1e^{-4}$ and we decay it at the 6-th epoch with a rate of 0.1.

4.3. Comparison with the State-of-the-arts

Quantitative comparison. We first compare our proposed method with the state-of-the-art methods on the GAICD [57] dataset in Table 1. Note that CGS [23] is trained on the conference version [56] of the GAICD

Model	$Acc_{1/5}$	$Acc_{2/5}$	$Acc_{3/5}$	$Acc_{4/5}$	$\overline{Acc_5}$	$Acc_{1/10}$	$Acc_{2/10}$	$Acc_{3/10}$	$Acc_{4/10}$	$\overline{Acc_{10}}$	$\overline{SRCC}$	$\overline{PCC}$
A2RL [21]	23.2	-	-	-	-	39.5	-	-	-	-	-	-
VPN [52]	36.0	-	-	-	-	48.5	-	-	-	-	-	-
VFN [5]	26.6	26.5	26.7	25.7	26.4	40.6	40.2	40.3	39.3	40.1	0.485	0.503
VEN [52]	37.5	35.0	35.3	34.2	35.5	50.5	49.2	48.4	46.4	48.6	0.616	0.662
GAIC [57]	68.2	64.3	61.3	58.5	63.1	84.4	82.7	80.7	78.7	81.6	0.849	0.874
CGS [23]	63.0	62.3	58.8	54.9	59.7	81.5	79.5	77.0	73.3	77.8	0.795	-
CGS* [23]	66.2	63.0	59.6	56.5	61.3	84.4	81.4	78.9	76.9	80.4	0.850	0.874
TransView [36]	69.0	**66.9**	61.9	57.8	63.9	85.4	84.1	81.3	78.6	82.4	0.857	0.880
Ours (w/o te)	68.4	65.1	62.1	59.2	63.7	86.2	83.1	81.4	79.5	82.6	0.865	0.889
Ours	**70.0**	**66.9**	**62.5**	**59.8**	**64.8**	**86.8**	**84.5**	**82.9**	**79.8**	**83.3**	**0.872**	**0.893**

Table 1. Quantitative comparison to other state-of-art methods on the GAICD dataset [57]. The best performance is in boldface. The line of CGS is reported on a part of the GAICD dataset [56] from paper [23], and CGS* is implemented by ourselves on the whole GAICD dataset [57]. The results of GAIC are copied from [57] and other methods are from [36].

Method	Training Set	IoU ↑	Disp ↓
A2RL [21]	AVA	0.663	0.089
A3RL [22]	AVA	0.696	0.077
VPN [52]	CPC	0.711	0.073
VEN [52]	CPC	0.735	0.072
ASM [46]	CPC	0.749	0.068
GAIC [57]	GAICD	0.672	0.084
CGS [23]	GAICD	0.685	0.079
TransView [36]	GAICD	0.682	0.080
Ours (w/o te)	GAICD	0.686	0.078
Ours	GAICD	0.695	0.075

Table 2. Quantitative comparison to other state-of-art methods on the FCDB dataset [4]. Note that previous works report the results using different training sets(AVA [32], CPC [52], GAICD [57]).

dataset. We report the results of CGS trained on the whole GAICD dataset of the journal version [57] as CGS* for fair comparison. The results of GAIC are copied from [57] and other methods are from [36].

We report two versions of our method, in which Ours (w/o te) does not use unlabeled test images while Ours is the full method in the transductive learning setting. We observe that our proposed model performs favorably against state-of-the-art methods on the GAICD dataset. Moreover, as our model uses the same backbone and region feature acquired by RoI&RoD Align as GAIC [57] and TransView [36], the comparison with [36, 57] shows the superiority of our proposed spatial-aware feature and rank consistency.

We also report the results of our proposed model on FCDB [4] dataset in Table 2. Note that previous works used different backbones and training sets. Compared with GAIC and TransView using the same backbone and training set as us, our model also achieves better performance.

Model complexity and runtime. We report the model complexity and runtime of VFN [5], VEN [52], VPN [52], CGS [23], GAIC [57], and our model in Table 3. Ours(basic) uses the same network as GAIC [57] but different in some implementation details (*e.g.*, learning rate decay, weight decay) and Ours is the entire network proposed. Note that, all models are run on the PC with Intel(R) Core(TM) i7-9700K CPU and one single NVIDIA GTX 1080Ti GPU. We can see the inference speed of our model is at the same level as GAIC and CGS, and much faster than VFN, VEN, and VPN. As we employ a light-weighted Encoder E_s to model the relationship between crops and aesthetic elements, the number of our model parameters and runtime are slightly increased compared with GAIC. However, it's still acceptable for mobile device applications considering cropping performance.

Method	Backbone	#Parameters	Runtime
VFN	Alexnet	14.88M	2491ms
VEN	VGG16	40.93M	5331ms
VPN	VGG16	65.31M	149ms
CGS	VGG16	21.25M	31ms
GAIC	MobileNetv2	5.91M	24ms
Ours(basic)	MobileNetv2	5.91M	25ms
Ours	MobileNetv2	7.10M	32ms

Table 3. Model complexity and runtime comparison. We report our proposed model and existing methods including VFN [5], VEN [52], VPN [52], CGS [23], and GAIC [57]. The runtime is the time to infer one image on average.

Qualitative comparison. In Figure 3, we provide a qualitative comparison with existing methods including VFN [5], VEN [52], VPN [52], CGS [23], and GAIC [57]. Only top-1 crops are shown for comparison among about 85 pre-defined anchors in an image. We can observe that important edges and salient targets appear at more appropriate locations in the crops obtained by our method, so that our crops own higher aesthetic quality and more appealing

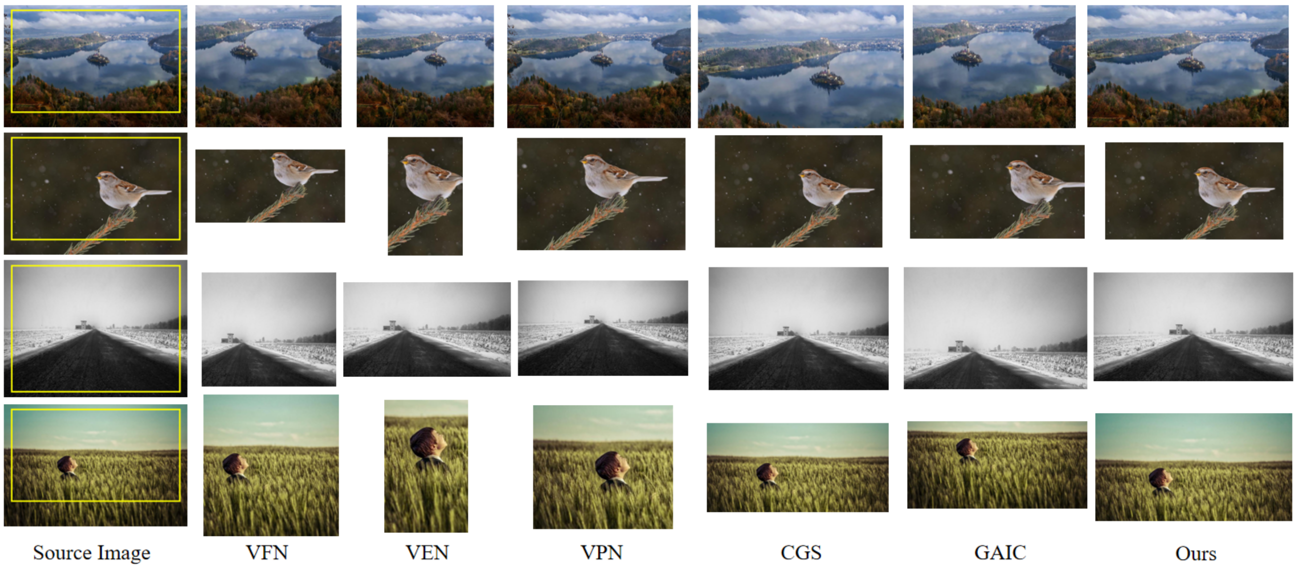

Figure 3. Qualitative comparison on GAICD test set. We show the annotated best crop (yellow bounding box) in the source image in the left column and top-1 crops obtained by different methods in the rest of the columns.

visual effect. For example, in the first row, our method preserves the entire shoreline of the lake without cutting through it, which makes the overall composition more appealing. In the last row, our method places the child at a better location obeying 'Rule of Thirds' and thus the crop has higher composition quality. Among those methods, CGS [23] is competitive probably due to its mutual relations modeling between crops. However, it lacks the ability to analyze and handle object edges and lines compared with our model. For example, in the first row, it cuts through the shoreline of the lake. More qualitative comparisons can be seen in Supplementary.

User study. The task of image cropping is subjective to a certain degree. Following previous works [23, 52, 58], we also conduct user study for more comprehensive comparison. We randomly sample 200 test images from GAICD [57] and FCDB [4] with a ratio of 1:1. For each test image, we collect the top-1 crops obtained by VFN [5], VEN [52], VPN [52], CGS [23], GAIC [57] and our method, and invite 20 annotators to choose the best crop from the six results. We count the ratio that each method is chosen as the best one, and the results are 7.6%, 10.7%, 6.5%, 24.5%, 18.9%, and 31.8% respectively corresponding to the methods abovementioned, showing that our model significantly outperforms other models.

4.4. Ablation study

In this section, we design three groups of ablation studies to explore the contribution of each component. All the ablation studies are conducted on the GAICD dataset [57].

Model components. Firstly, we investigate the impact of each component in our model. We set our basic network using only region features F_r to predict aesthetic scores that is similar to GAIC [56, 57]. Then, we add the spatial-aware feature and rank consistency components respectively, and finally use both of them. The results are shown in Table 4. The difference between row 1 in Table 4 and the GAIC result in Table 1 is caused by the implementation details. We can draw the following conclusions: a) when only using the

Row	SAF	RC	$\overline{SRCC}\uparrow$	$\overline{PCC}\uparrow$	$\overline{Acc_5}\uparrow$	$\overline{Acc_{10}}\uparrow$
1			0.858	0.882	61.7	80.5
2	✓		0.865	0.889	63.7	82.6
3		✓	0.868	0.890	64.0	82.3
4	✓	✓	**0.872**	**0.893**	**64.8**	**83.3**

Table 4. Ablation study of different components in our model. "SAF" and "RC" are short for Spatial-aware Feature and Rank Consistency respectively.

Row	Agg	Att	CM	$\overline{SRCC}\uparrow$	$\overline{PCC}\uparrow$	$\overline{Acc_5}\uparrow$	$\overline{Acc_{10}}\uparrow$
1	Cat		✓	0.863	0.886	63.0	81.9
2	Avg		✓	0.862	0.885	62.9	81.7
3	Cat	✓	✓	**0.865**	**0.889**	**63.7**	**82.6**
4	Cat	✓		0.860	0.883	62.5	82.3

Table 5. Ablation study of the spatial-aware feature. "Agg", "Att" and "CM" are short for Aggregate, Attention and Crop Mask respectively. "Cat" and "Avg" are short for concatenation and average respectively, which are two ways to aggregate multiple layers of feature maps. "Attention" means whether using channel attention. "Crop mask" means whether appending the crop mask.

spatial-aware feature or rank consistency component, correlation coefficient metrics and return K of top-N accuracy are improved, which proves that both the spatial-aware feature or rank consistency are effective; b) When we train our model jointly using the spatial-aware feature and rank consistency, the performance is further improved, which implies that these two components are complementary.

In order to gain an intuition on how each component improves cropping results, we show some examples of GAICD [57] test set using the basic network and our proposed method with only spatial-aware feature component and rank consistency component respectively in Supplementary.

Channel attention. To figure out how the channel attention block behaves in the spatial-aware feature, we conduct this group of ablation studies. We investigate how to aggregate multiple layers of feature maps, whether to use channel attention, and whether to append the crop mask. Note that we do not use rank consistency in this subsection. The results are shown in Table 5. The comparison between row 1 and row 2 shows that concatenation works slightly better than average. The comparison between row 3 and row 1 verifies the effectiveness of channel attention. The comparison between row 4 and row 3 verifies the necessity of appending crop masks to the aggregated feature map, which allows the model to capture the spatial relationship between each crop and aesthetic elements.

To better understand the working mechanism of channel attention, we visualize each channel (layer of feature map) with their attention value in Figure 4. We observe that the channel attention distribution becomes stable when the training process converges. The attention values of some channels are higher than others, which implies that the model learns to emphasize or suppress certain channels automatically. As shown in Figure 4, layers of 4,7,9 are suppressed, while layers of 2,5,11,14 are emphasized. Intuitively, we can see that the feature maps with high attention values exhibit clearer edges and more notable salient objects, which are helpful to model the spatial relationship between candidate crops and aesthetic elements.

The above two groups of ablation studies prove that our proposed spatial-aware feature can capture the spatial relationship between candidate crops and aesthetic elements indeed when using different layers of feature maps and crop masks properly. Furthermore, the model can learn how to select and aggregate information from different layers automatically. Therefore, our model can be aware of which aesthetic elements should be included and where they should be placed, leading to visually appealing crops.

Ranking knowledge transfer. We further conduct ablation studies on rank consistency which prove that ranking knowledge is transferrable from labeled images to unlabeled images. We also compare our ranking consistency method with other alternative approaches same as [5, 52] for transductive learning. The results show that two alternative ways to use unlabeled test images cannot exceed our proposed rank consistency. The details about the performance of the pair-wise ranking classifier and the comparison between our ranking consistency method and other alternative approaches to use unlabeled data are in Supplementary.

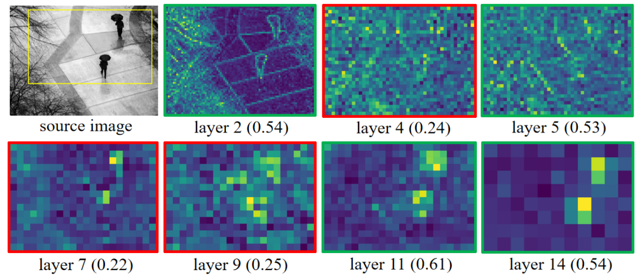

Figure 4. Channel attention visualization. We select 3 channels (in red boxes) with the lowest attention values and 4 channels (in green boxes) with the highest attention values and show their corresponding layer of feature map and attention values.

5. Limitations

Although our method can generally achieve satisfactory results, there still exist some failure cases. When cropping landscape photos, our model usually performs better than other approaches, because it tends to crop a broad view that contains salient objects as many as possible and place them with good composition quality. However, when some overlength edges cross half of the image, the crops may preserve those edges and the holistic composition quality is compromised. The visualization results could be found in Supplementary.

6. Conclusion

In this work, we have proposed a novel spatial-aware feature to capture the spatial relationship between candidate crops and aesthetic elements. We have also proposed to transfer ranking knowledge from labeled images to unlabeled images and enforce ranking consistency on unlabeled images. Quantitative and qualitative comparisons have shown that our method obtains the state-of-the-art performance on benchmark datasets.

Acknowledgement

The work was supported by the National Science Foundation of China (62076162), and the Shanghai Municipal Science and Technology Major/Key Project, China (2021SHZDZX0102, 20511100300).

References

[1] Andrew Arnold, Ramesh Nallapati, and William W Cohen. A comparative study of methods for transductive transfer learning. In *ICDMW*, 2007. 3, 4

[2] Jiansheng Chen, Gaocheng Bai, Shaoheng Liang, and Zhengqin Li. Automatic image cropping: A computational complexity study. In *CVPR*, 2016. 1, 2

[3] Li-Qun Chen, Xing Xie, Xin Fan, Wei-Ying Ma, Hong-Jiang Zhang, and He-Qin Zhou. A visual attention model for adapting images on small displays. *Multimedia systems*, 9(4):353–364, 2003. 1, 2

[4] Yi-Ling Chen, Tzu-Wei Huang, Kai-Han Chang, Yu-Chen Tsai, Hwann-Tzong Chen, and Bing-Yu Chen. Quantitative analysis of automatic image cropping algorithms: A dataset and comparative study. In *WACV*, 2017. 1, 2, 5, 6, 7

[5] Yi-Ling Chen, Jan Klopp, Min Sun, Shao-Yi Chien, and Kwan-Liu Ma. Learning to compose with professional photographs on the web. In *ACMMM*, 2017. 1, 2, 3, 6, 7, 8

[6] Bin Cheng, Bingbing Ni, Shuicheng Yan, and Qi Tian. Learning to photograph. In *ACMMM*, 2010. 1, 2

[7] Yang Cheng, Qian Lin, and Jan P Allebach. Re-compose the image by evaluating the crop on more than just a score. In *WACV*, 2022. 1, 2

[8] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Studying aesthetics in photographic images using a computational approach. In *ECCV*, 2006. 2

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 3

[10] Yubin Deng, Chen Change Loy, and Xiaoou Tang. Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine*, 34(4):80–106, 2017. 1

[11] Yubin Deng, Chen Change Loy, and Xiaoou Tang. Aesthetic-driven image enhancement by adversarial learning. In *ACMMM*, 2018. 2

[12] Chen Fang, Zhe Lin, Radomir Mech, and Xiaohui Shen. Automatic image cropping using visual composition, boundary simplicity and content preservation models. In *ACMMM*, 2014. 1, 2

[13] Guanjun Guo, Hanzi Wang, Chunhua Shen, Yan Yan, and Hong-Yuan Mark Liao. Automatic image cropping for visual aesthetic enhancement using deep neural networks and cascaded regression. *IEEE Transactions on Multimedia*, 20(8):2073–2085, 2018. 2

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 2, 3

[15] Chaoyi Hong, Shuaiyuan Du, Ke Xian, Hao Lu, Zhiguo Cao, and Weicai Zhong. Composing photos like a photographer. In *CVPR*, 2021. 2

[16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 2, 4

[17] Gengyun Jia, Huaibo Huang, Chaoyou Fu, and Ran He. Rethinking image cropping: Exploring diverse compositions from global views. In *CVPR*, 2022. 2

[18] Yueying Kao, Ran He, and Kaiqi Huang. Automatic image cropping with aesthetic map and gradient energy map. In *ICASSP*, 2017. 2

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 5

[20] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *ICLR*, 2017. 2

[21] Debang Li, Huikai Wu, Junge Zhang, and Kaiqi Huang. A2-rl: Aesthetics aware reinforcement learning for image cropping. In *CVPR*, 2018. 1, 2, 6

[22] Debang Li, Huikai Wu, Junge Zhang, and Kaiqi Huang. Fast a3rl: Aesthetics-aware adversarial reinforcement learning for image cropping. *IEEE Transactions on Image Processing*, 28(10):5105–5120, 2019. 2, 6

[23] Debang Li, Junge Zhang, Kaiqi Huang, and Ming-Hsuan Yang. Composing good shots by exploiting mutual relations. In *CVPR*, 2020. 1, 2, 5, 6, 7

[24] Xuewei Li, Xueming Li, Gang Zhang, and Xianlin Zhang. Image aesthetic assessment using a saliency symbiosis network. *Journal of Electronic Imaging*, 28(2):023008, 2019. 2

[25] Zhuopeng Li and Xiaoyan Zhang. Collaborative deep reinforcement learning for image cropping. In *ICME*, 2019. 2

[26] Ligang Liu, Renjie Chen, Lior Wolf, and Daniel Cohen-Or. Optimizing photo composition. In *Computer graphics forum*, volume 29, pages 469–478, 2010. 1, 2

[27] Peng Lu, Hao Zhang, Xujun Peng, and Xiaofu Jin. An end-to-end neural network for image cropping by learning composition from aesthetic photos. *arXiv preprint arXiv:1907.01432*, 2019. 2

[28] Weirui Lu, Xiaofen Xing, Bolun Cai, and Xiangmin Xu. Listwise view ranking for image cropping. *IEEE Access*, 7:91904–91911, 2019. 2

[29] Matthew Ma and Jinhong K Guo. Automatic image cropping for mobile device with built-in camera. In *CCNC*, 2004. 2

[30] Long Mai, Hailin Jin, and Feng Liu. Composition-preserving deep photo aesthetics assessment. In *CVPR*, 2016. 2

[31] Luca Marchesotti, Claudio Cifarelli, and Gabriela Csurka. A framework for visual saliency detection with applications to image thumbnailing. In *ICCV*, 2009. 1, 2

[32] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *CVPR*, 2012. 6

[33] Masashi Nishiyama, Takahiro Okabe, Yoichi Sato, and Imari Sato. Sensation-based photo cropping. In *ACMMM*, 2009. 1, 2

[34] Yassine Ouali, Céline Hudelot, and Myriam Tami. An overview of deep semi-supervised learning. *arXiv preprint arXiv:2006.05278*, 2020. 2, 3

[35] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009. 3

[36] Zhiyu Pan, Zhiguo Cao, Kewei Wang, Hao Lu, and Weicai Zhong. Transview: Inside, outside, and across the cropping view boundaries. In *ICCV*, 2021. 1, 2, 5, 6

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 4

[38] Ellen Riloff. Automatically generating extraction patterns from untagged text. In *AAAI*, 1996. 2

[39] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018. 1, 3, 5

[40] Anthony Santella, Maneesh Agrawala, Doug DeCarlo, David Salesin, and Michael Cohen. Gaze-based interaction for semi-automatic photo cropping. In *ACM SIGCHI*, pages 771–780, 2006. 2

[41] Henry Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371, 1965. 2

[42] Fred Stentiford. Attention based auto image cropping. In *ICCV*, 2007. 2

[43] Bongwon Suh, Haibin Ling, Benjamin B Bederson, and David W Jacobs. Automatic thumbnail cropping and its effectiveness. In *UIST*, 2003. 1, 2

[44] Jin Sun and Haibin Ling. Scale and object aware image thumbnailing. *International journal of computer vision*, 104(2):135–153, 2013. 1, 2

[45] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 2

[46] Yi Tu, Li Niu, Weijie Zhao, Dawei Cheng, and Liqing Zhang. Image cropping with composition and saliency aware aesthetic score map. In *AAAI*, 2020. 1, 2, 6

[47] Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020. 2, 3

[48] Vladimir N Vapnik. Statistical learning theory. *Adaptive and learning systems for signal processing communications and control*, 1998. 3, 4

[49] Eleonora Vig, Michael Dorr, and David Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *CVPR*, 2014. 1, 2

[50] Wenguan Wang and Jianbing Shen. Deep cropping via attention box prediction and aesthetics assessment. In *ICCV*, 2017. 2

[51] Wenguan Wang, Jianbing Shen, and Haibin Ling. A deep network solution for attention and aesthetics aware photo cropping. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1531–1544, 2018. 2

[52] Zijun Wei, Jianming Zhang, Xiaohui Shen, Zhe Lin, Radomir Mech, Minh Hoai, and Dimitris Samaras. Good view hunting: Learning photo composition from dense view pairs. In *CVPR*, 2018. 1, 2, 3, 6, 7, 8

[53] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268, 2020. 2

[54] Jianzhou Yan, Stephen Lin, Sing Bing Kang, and Xiaoou Tang. Learning the change for automatic image cropping. In *CVPR*, 2013. 1, 2, 4

[55] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *ACL*, 1995. 2

[56] Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Reliable and efficient image cropping: A grid anchor based approach. In *CVPR*, 2019. 1, 2, 3, 4, 5, 6, 7

[57] Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Grid anchor based image cropping: A new benchmark and an efficient model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[58] Bo Zhang, Li Niu, Xing Zhao, and Liqing Zhang. Human-centric image cropping with partition-aware and content-preserving features. *ECCV*, 2022. 2, 7

[59] Jianming Zhang and Stan Sclaroff. Saliency detection: A boolean map approach. In *ICCV*, 2013. 1, 2

[60] Luming Zhang, Mingli Song, Yi Yang, Qi Zhao, Chen Zhao, and Nicu Sebe. Weakly supervised photo cropping. *IEEE Transactions on Multimedia*, 16(1):94–107, 2013. 1, 2

[61] Luming Zhang, Mingli Song, Qi Zhao, Xiao Liu, Jiajun Bu, and Chun Chen. Probabilistic graphlet transfer for photo cropping. *IEEE Transactions on Image Processing*, 22(2):802–815, 2012. 2

[62] Mingju Zhang, Lei Zhang, Yanfeng Sun, Lin Feng, and Weiying Ma. Auto cropping for digital photographs. In *ICME*, 2005. 1, 2

[63] Lei Zhong, Feng-Heng Li, Hao-Zhi Huang, Yong Zhang, Shao-Ping Lu, and Jue Wang. Aesthetic-guided outward image cropping. *ACM Transactions on Graphics (TOG)*, 40(6):1–13, 2021. 1, 2