Human-centric Image Cropping with Partition-aware and Content-preserving Features

Bo Zhang, Li Niu[*], Xing Zhao, and Liqing Zhang

MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University, China
{bo-zhang,ustcnewly,1033874657}@sjtu.edu.cn, zhang-lq@cs.sjtu.edu.cn

Abstract. Image cropping aims to find visually appealing crops in an image, which is an important yet challenging task. In this paper, we consider a specific and practical application: human-centric image cropping, which focuses on the depiction of a person. To this end, we propose a human-centric image cropping method with two novel feature designs for the candidate crop: partition-aware feature and content-preserving feature. For partition-aware feature, we divide the whole image into nine partitions based on the human bounding box and treat different partitions in a candidate crop differently conditioned on the human information. For content-preserving feature, we predict a heatmap indicating the important content to be included in a good crop, and extract the geometric relation between the heatmap and a candidate crop. Extensive experiments demonstrate that our method can perform favorably against state-of-the-art image cropping methods on human-centric image cropping task. Code is available at https://github.com/bcmi/Human-Centric-Image-Cropping.

1 Introduction

Image cropping aims to automatically find visually appealing crops in an image, which is critical in various down-stream applications, *e.g.*, photo post-processing [6], view recommendation [21,41,20], image thumbnailing [10,3], and camera view adjustment suggestion [35]. In this paper, we address a specific and practical application: human-centric image cropping, which focuses on the depiction of a person and benefits a variety of applications, including portrait enhancement [47] and portrait composition assistance [48,49]. For a human-centric image, a good crop depends on the position of the human in the crop, human information, and the content of interest, which makes human-centric image cropping challenging.

Several previous works [48,2,47] have already focused on portrait photograph cropping, which extracted hand-crafted features from the results of saliency detection, human face detection, or human pose estimation. However, extracting hand-crafted features is laborious and the hand-crafted features are generally not robust for modeling the huge aesthetic space [9]. Recently, numerous methods [15,44,37,23,22,40] addressed image cropping task in a data-driven manner, in

[*] Corresponding author

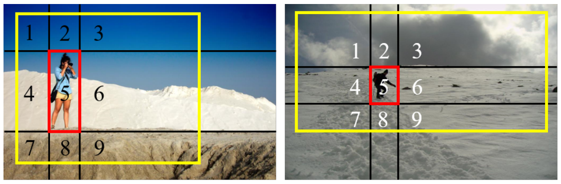

Fig. 1. Illustration of the motivation behind partition-aware feature. The whole image is divided into nine partitions based on the human bounding box (red). To produce the best crop (yellow), the aesthetic contribution of similar content in different partitions depends on its relative position to human subject

which models are directly trained with the human-annotated datasets [5,43,41]. However, for human-centric images, these methods rarely explicitly consider human information. In contrast, we show that exploiting human information can significantly help obtain good crops. Based on the general pipeline of data-driven methods, we propose two innovations for human-centric image cropping.

In this paper, we refer to the images that meet the following conditions as human-centric images: 1) The image subject is single person, while there can be other people in the background. 2) The area of the human bounding box does not exceed 90% of the entire frame. Given a human-centric image, the whole image can be divided into nine partitions based on the human bounding box (see Figure 1). Generally, the aesthetic contribution of similar content in different partitions depends on its relative position to the human subject. For example, in Figure 1, partitions 4 and 6 in the left subfigure have similar content, but the best crop preserves more content in partition 6 because the person looks to the right, making the content in partition 6 visually more important [12]. Similarly, partition 4 and partition 8 in the right subfigure also have similar content, but the best crop preserves more content in partition 4, probably because the person is moving forward and the content behind him becomes less important. Therefore, when extracting features of candidate crops for aesthetic evaluation, we should consider the partition location and human information (*e.g.*, human posture, face orientation). To this end, we propose a novel partition-aware feature by incorporating partition and human information, which enables treating different partitions in a candidate crop differently conditioned on the human information.

Furthermore, a good crop should preserve the important content of source image [11], which is dubbed as "content-preserving". However, to the best of our knowledge, there is no image cropping dataset that provides the annotation of important content. Existing methods [25,39,4,11] determine important content mainly based on their visual saliency by assuming that the most salient object is the most important content. In human-centric images, important content may imply key human parts (*e.g.*, face, hands), interesting objects (*e.g.*, landmark), and the objects (*e.g.*, racket, bicycle) that person interacts with. However, as shown in Figure 2, saliency may not capture the abovementioned objects very well. Here we adopt an unsupervised saliency detection method [13]. We have also

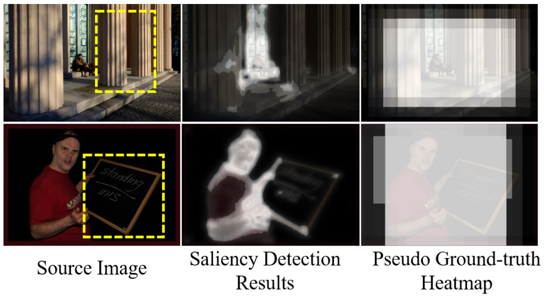

Fig. 2. The comparison between saliency detection [13] and the pseudo ground-truth heatmap of important content (see Section 3.3). The important content of human-centric images may contain interesting objects (*e.g.*, the landmark in the top row) and the objects that person interacts with (*e.g.*, the blackboard in the bottom row)

tried several supervised methods [16,50,7], which proves to be less effective. This is probably because that the unsupervised method has no dependence on training data and generalizes better on the image cropping datasets. Given an image with multiple annotated candidate crops [43,41], we conjecture that the candidate crops with relatively high scores are more likely to contain important objects. Thus, we use highly scored crops to produce pseudo ground-truth heatmap of important content (see Figure 2), which is used to supervise the heatmap prediction. Additionally, previous content-preserving methods [1,4] typically designed a hand-crafted algorithm based on certain principles (*e.g.*, maintaining the most salient region). Differently, we automatically learn content-preserving feature to capture the geometric relation between the predicted heatmap and each candidate crop, which represents how well each candidate crop preserves the important content.

Finally, for each candidate crop, we extract its partition-aware feature and content-preserving feature to predict an aesthetic score. The main contributions of this paper can be summarized as follows: 1) We propose a novel partition-aware feature to improve human-centric image cropping by exploiting human information, which allows to treat different regions in a candidate crop differently conditioned on the human information. 2) We design a novel approach to locate important content and a novel content-preserving feature to characterize the preservation of important content in a candidate crop. 3) We demonstrate that our model outperforms the state-of-the-art image cropping methods on the human-centric images of several benchmark datasets.

2 Related Work

Following [44,23], we divide existing image cropping methods into three categories according to the criteria for evaluating candidate crops, *i.e.*, attention-guided, aesthetics-informed, and data-driven.

Attention-guided Image Cropping: Attention-guided methods [27,25,26,4,11,36,47] assumed that the best crops should preserve visually important content, which is usually determined by the saliency detection methods [38,17]. Usually, the view with the highest average saliency score is selected as the best crop. However, saliency may not accurately reflect the content of interest for human-centric images (see Figure 2). Differently, we assume that the content that appears in multiple highly scored crops is more likely to be important content, leading to more flexible and practical important content estimation.
Aesthetics-informed Image Cropping: The aesthetics-informed methods evaluated candidates by comparing the overall aesthetic quality of different crops. To achieve this, earlier methods [45,42,46] usually employed hand-crafted features or composition rules. However, the simple hand-crafted features may not accurately predict the complicated image aesthetics [44].
Data-driven Image Cropping: Most recent methods address the task in a data-driven manner. Some methods [6,18,29] trained a general aesthetic evaluator on image aesthetic datasets to facilitate image cropping. With the aid of image cropping datasets [5,43,41], numerous methods [15,43,37,28,23,41] used pairwise learning to train an end-to-end model on these datasets, which can generate crop-level scores for ranking different candidate crops.

Our method is developed based on the general pipeline of the data-driven methods, but is specially tailored to human-centric image cropping with two innovations, *i.e.*, partition-aware and content-preserving features.

3 Methodology

3.1 Overview

The flowchart of the proposed method is illustrated in Figure 3, in which we adopt a similar pipeline as [44,23]. Given an image, we first integrate multi-scale feature maps from a pretrained backbone (*e.g.*, VGG16 [34]) to obtain the basic feature map. After that, we update the basic feature map to the partition-aware feature map, based on which we extract partition-aware region feature and content-preserving feature for each candidate crop. Finally, we predict the crop-level score using the concatenation of partition-aware region feature and content-preserving feature.

3.2 Partition-aware Feature

To acquire the human bounding box, we leverage Faster R-CNN [31] trained on Visual Genome [19] to detect human subjects for human-centric images. We check each image to ensure that the predicted bounding box correctly encloses the main human subject. We describe how to determine the main subject and discuss the robustness of our method against human detection in Supplementary. As illustrated in Figure 1, the whole image can be divided into nine non-overlapping partitions by the human bounding box. We conjecture that the aesthetic value

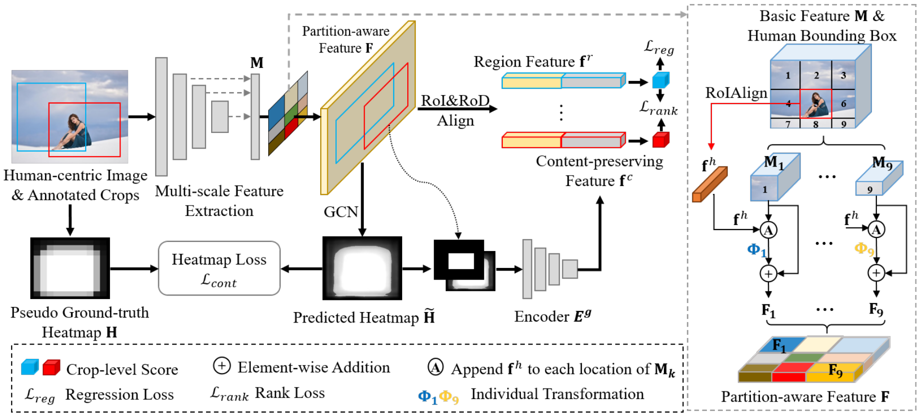

Fig. 3. The flowchart of our method for human-centric image cropping (left) and the proposed partition-aware feature (right). We use pretrained VGG16 [34] as backbone to extract basic feature map $\mathbf{M}$, from which we derive partition-aware feature map $\mathbf{F}$. Besides, we use the region feature obtained by RoIAlign [14] and RoDAlign [43] schemes, and content-preserving feature to predict scores for each candidate crop

of similar content in different partitions often varies with its relative position to the human subject, so the feature map should be partition-aware.

To achieve this goal, we derive partition-aware feature map from the basic feature map, as illustrated in the left subfigure in Figure 3. Given an $H \times W$ basic feature map $\mathbf{M}$ with C channels, we partition it into nine regions $\{\mathbf{M}_k|_{k=1}^{9}\}$ with $\mathbf{M}_k$ being the k-th partition. Considering that the relative position of each partition to human subject is conditioned on the human information (*e.g.*, face orientation and posture), as exemplified in Figure 1, we extract human feature $\mathbf{f}^h$ using RoIAlign [14] to implicitly encode the aforementioned human information and reduce its dimension to $\frac{C}{2}$, which is appended to each location of the basic feature map. The resultant feature map is represented by $\hat{\mathbf{M}}$ with $\hat{\mathbf{M}}_k$ being the k-th partition. To explicitly tackle different partitions differently, we employ nine individual nonlinear transformations for each partition to update their features with residual learning by

$$\mathbf{F}_k = \Phi_k(\hat{\mathbf{M}}_k) + \mathbf{M}_k, \tag{1}$$

where we use a 3×3 convolutional (conv) layer with C output channels followed by ReLU as the transformation function $\Phi_k(\cdot)$.

After that, we obtain a new feature map $\mathbf{F}$ by combining all updated partitions $\{\mathbf{F}_k|_{k=1}^{9}\}$, which has the same size as the basic feature map $\mathbf{M}$. *By integrating human feature and employing partition-specific transformations, similar contents in different partitions can produce different responses conditioned on the human information.* Thus, we refer to $\mathbf{F}$ as partition-aware feature map. Following [43,23], given a candidate crop, we employ RoIAlign [14] and RoDAlign [43] schemes to extract its partition-aware region feature based on $\mathbf{F}$, denoted as $\mathbf{f}^r$.

3.3 Content-preserving Feature

Apart from the position of the human in the crop and human information, the preservation of important content also plays an important role when evaluating candidate crops. So we propose to predict a heatmap to indicate the location of important content and then automatically learn content-preserving features to augment our method.

Graph-based Region Relation Mining: Considering that important content estimation may benefit from exploiting the mutual relation between different regions, we construct a graph over the partition-aware feature map $\mathbf{F}$ and apply graph convolution [32,24]. Specifically, we reshape the partition-aware feature maps into a matrix $\bar{\mathbf{F}} \in \mathcal{R}^{L \times C}$, where $L = H \times W$. Each pixel-wise feature vector in $\bar{\mathbf{F}}$ is a graph node that represents one local region in the image.

To model the relation between pairwise regions, we define the adjacency matrix $\mathbf{A} \in \mathcal{R}^{L \times L}$ according to the cosine similarity of region features following [33]. Then we perform reasoning over the graph $\bar{\mathbf{F}}$ by graph convolution [32]:

$$\bar{\mathbf{F}}' = \sigma(\mathbf{A}\bar{\mathbf{F}}\mathbf{\Theta}), \tag{2}$$

where $\mathbf{\Theta} \in \mathcal{R}^{C \times C}$ is the the trainable weight matrix of the graph convolution layer and $\sigma(\cdot)$ is ReLU activation. Then, we reshape $\bar{\mathbf{F}}'$ back to $\mathbf{F}' \in \mathcal{R}^{H \times W \times C}$. Compared with the conventional convolution, graph convolution allows the message flow across local regions, which is helpful for important content prediction (see Section 4.6).

Important Content Estimation: To obtain a high-resolution feature map for fine-grained important content localization, we upsample $\mathbf{F}'$ by four times followed by 3×3 conv and ReLU. Based on the upsampled feature map, we apply a prediction head (*i.e.*, 1×1 conv followed by Sigmoid function) to produce a heatmap $\widetilde{\mathbf{H}}$ in the range of $[0, 1]$, in which larger score indicates more important content.

There is no ground-truth heatmap for important content. Nevertheless, existing cropping datasets [43,41] are associated with multiple scored crops for each image, with a larger score indicating higher aesthetic quality. *Based on the assumption that highly scored crops are more likely to contain important content, we propose to generate pseudo ground-truth heatmap from the weighted average of highly scored crops.*

Specifically, given an image with multiple candidate crops, we suppose the score of the m-th crop to be y_m. We take the average score of all crops in the dataset as the threshold to select highly scored crops, and convert the bounding box of each selected crop to a binary map, which is resized to the same size as the predicted heatmap $\widetilde{\mathbf{H}}$. We obtain the pseudo ground-truth heatmap $\mathbf{H}$ via the weighted average of all binary maps, in which larger weight is assigned to the crop with higher score. Here, we perform softmax normalization to produce the weights for each highly scored crop: $\omega_m = \frac{\exp(y_m)}{\sum_{m=1}^{N_h} \exp(y_m)}$, in which N_h is the number of highly scored crops. In the training stage, we employ an L1 loss to supervise the heatmap learning:

$$\mathcal{L}_{cont} = \|\mathbf{H} - \widetilde{\mathbf{H}}\|_1. \tag{3}$$

As demonstrated in Figure 2, the pseudo ground-truth heatmaps can highlight the attractive objects in the background and the objects that person interacts with, which are often ignored by saliency detection methods.

Content-preserving Feature Extraction: To leverage the prior information of important content [41], previous content-preserving methods typically relied on certain heuristic principles (*e.g.*, minimizing the cropping area while maximizing its attention value [4]), which require extensive manual designs. *Instead, we guide the network to automatically learn how well each candidate crop preserves the important content. The idea is learning a content-preserving feature to capture the geometric relation between the heatmap and each candidate crop.* Specifically, for the m-th candidate crop, we concatenate its corresponding binary map and the predicted heatmap $\widetilde{\mathbf{H}}$ channel-wisely. Then, we apply an encoder E^g to the concatenated maps to extract the content-preserving feature $\mathbf{f}^c$.

Finally, for each candidate crop, we concatenate its partition-aware region feature $\mathbf{f}^r$ and content-preserving feature $\mathbf{f}^c$, which is passed through a fully connected (fc) layer to get the aesthetic score for this crop.

3.4 Network Optimization

We train the proposed model with a multi-task loss function in an end-to-end manner. Given an image containing N candidate crops, the ground-truth and predicted scores of the m-th crop are denoted by y_m and $\tilde{y}_m$, respectively. We first employ a smooth L1 loss [31] for the score regression considering its robustness to outliers:

$$\mathcal{L}_{reg} = \frac{1}{N}\sum_{m=1}^{N} \mathcal{L}_{s1}(y_m - \tilde{y}_m), \tag{4}$$

where $\mathcal{L}_{s1}(\cdot)$ represents the smooth L1 loss:

$$\mathcal{L}_{s1}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1, \\ |x| - 0.5, & \text{otherwise.} \end{cases} \tag{5}$$

Besides regression loss, we also use a ranking loss [23] to learn the relative ranking order between pairwise crops explicitly, which is beneficial for enhancing the ability of ranking crops with similar content. With $e_{m,n} = y_m - y_n$ and $\tilde{e}_{m,n} = \tilde{y}_m - \tilde{y}_n$, the ranking loss is computed by

$$\mathcal{L}_{rank} = \frac{\sum_{m,n} \max\big(0, \mathrm{sign}(e_{m,n})(e_{m,n} - \tilde{e}_{m,n})\big)}{N(N-1)/2}. \tag{6}$$

After including the heatmap prediction loss in Eqn.(3), the total loss is summarized as

$$\mathcal{L} = \mathcal{L}_{reg} + \mathcal{L}_{rank} + \lambda\mathcal{L}_{cont}, \tag{7}$$

in which the trade-off parameter λ is set as 1 via cross-validation (see Supplementary).

Limited by the small number of annotated human-centric images in existing image cropping datasets [43,41], only training on human-centric images would lead to weak generalization ability to the test set. Therefore, *we employ both human-centric and non-human-centric images to train our model.* For non-human-centric images, we use the basic feature map $\mathbf{M}$ to replace the partition-aware feature map $\mathbf{F}$, because there could be no dominant subject used to partition the image. Besides, we extract content-preserving features from non-human-centric images in the same way as human-centric images, because preserving important content is also crucial for non-human-centric images. In this way, our model is able to train and infer on both human-centric images and non-human-centric images.

4 Experiments

4.1 Datasets and Evaluation Metrics

We conduct experiments on the recent GAICD dataset [43], which contains 1,236 images (1,036 for training and 200 for testing). Each image has an average of 86 annotated crops. There are 339 and 50 human-centric images in the training set and test set of GAICD dataset, respectively. As described in Section 3.4, we employ the whole training set (1,036 images) for training and evaluate on the human-centric samples of the test set. Following [43], three evaluation metrics are employed in our experiments, including the average Spearman's rank-order correlation coefficient ($\overline{SRCC}$) and averaged top-N accuracy ($\overline{Acc_N}$) for both $N = 5$ and $N = 10$. The $\overline{SRCC}$ computes the rank correlation between the ground-truth and predicted scores of crops for each image, which is used to evaluate the ability of correctly ranking multiple annotated crops. The $\overline{Acc_N}$ measures the ability to return the best crops.

Apart from GAICD dataset, we also collect 176 and 39 human-centric images from existing FCDB [5] and FLMS [11] datasets, respectively. We evaluate our method on the collected 215 human-centric images from two datasets, following the experimental setting in [28,37]: training the model on CPC dataset [41], and using intersection of union (IoU) and boundary displacement (Disp) for performance evaluation. Note that we train on the whole CPC dataset, which contains 10,797 images including 1,154 human-centric images and each image has 24 annotated crops.

4.2 Implementation Details

Following existing methods [15,37,23], we use VGG16 [34] pretrained on ImageNet [8] as the backbone. We apply 1×1 conv to unify the channel dimensions of the last three output feature maps as 256 and add up three feature maps to produce the multi-scale feature map. We reduce the channel dimension of multi-scale feature map to 32 using a 1×1 conv, *i.e.*, $C = 32$. E^g is implemented by

Table 1. Ablation studies of the proposed method. $\mathbf{f}^h$: human feature. "res": update partition-aware feature with residual learning. K: number of partitions. "conv": replace GCN with standard convolution. $\widetilde{\mathbf{H}}$: predicted important content heatmap. "saliency": replace heatmap with saliency map [13]. $\mathbf{f}^c$: content-preserving feature

	Partition	Content	$\overline{SRCC}\uparrow$	$\overline{Acc_5}\uparrow$	$\overline{Acc_{10}}\uparrow$
1			0.744	52.0	70.5
2	✓		0.774	54.8	74.3
3	w/o $\mathbf{f}^h$		0.769	54.2	73.8
4	w/o res		0.764	53.9	73.5
5	$K=1$		0.746	52.1	70.8
6	$K=2$		0.756	53.2	72.4
7		✓	0.781	56.8	75.6
8		conv	0.762	54.0	73.0
9		w/o $\widetilde{\mathbf{H}}$	0.741	50.9	69.5
10		saliency	0.752	52.5	71.8
11		only $\mathbf{f}^c$	0.643	35.2	49.1
12	✓	✓	**0.795**	**59.7**	**77.0**

two 3×3 convs and pooling operations followed by a fc layer. The dimensions of partition-aware region feature $\mathbf{f}^r$ and content-preserving feature $\mathbf{f}^c$ are both 256. Similar to [23,43], the short side of input images is resized to 256 and the aspect ratios remain unchanged. We implement our method using PyTorch [30] and set the random seed to 0. More implementation details can be found in Supplementary.

4.3 Ablation study

In this section, we start from the general pipeline of existing methods [44,23] and evaluate the effectiveness of two types of features. The results are summarized in Table 1. In the baseline (row 1), we only use the region feature extracted from the basic feature map $\mathbf{M}$ to predict scores for each crop.

Partition-aware Feature: Based on row 1, we replace the region feature with our proposed partition-aware region feature $\mathbf{f}^r$ in row 2, which verifies the effectiveness of partition-aware feature. Next, we conduct ablation studies based on row 2. First, we remove the human feature $\mathbf{f}^h$ and observe performance drop, which corroborates the importance of conditional human information. In Eqn.(1), we adopt $\mathit{\Phi}_k(\cdot)$ to learn the residual. Based on row 2, we remove the residual strategy by using $\mathbf{F}_k = \mathit{\Phi}_k(\hat{\mathbf{M}}_k)$. The comparison between row 4 and row 2 demonstrates the benefit of residual learning. Recall that each image is divided into nine partitions by the human bounding box. Based on row 2, we explore using one partition ($K = 1$) and two partitions ($K = 2$). When $K = 1$, we apply the same transformation $\mathit{\Phi}(\cdot)$ to the whole image. When $K = 2$, we divide the image into human bounding box and the outside region. By comparing $K = 1, 2, 9$,

Table 2. Comparison with the state-of-the-art methods on human-centric images in GAICD [43] dataset. GAIC(ext) [44] is the extension of GAIC[43]. The results marked with * are obtained using the released models from original papers

Method	Backbone	Training Data	$\overline{SRCC} \uparrow$	$\overline{Acc_5} \uparrow$	$\overline{Acc_{10}} \uparrow$
VFN* [6]	AlextNet	Flickr	0.332	10.1	21.1
VFN [6]	VGG16	GAICD	0.648	41.3	60.2
VEN* [41]	VGG16	CPC	0.641	22.4	36.2
VEN [41]	VGG16	GAICD	0.683	50.1	65.1
ASM-Net [37]	VGG16	GAICD	0.680	44.8	64.5
LVRN* [28]	VGG16	CPC	0.664	30.7	49.0
LVRN [28]	VGG16	GAICD	0.716	44.8	66.0
GAIC(ext)* [44]	MobileNetV2	GAICD	0.773	54.0	73.0
GAIC(ext) [44]	VGG16	GAICD	0.741	53.3	69.6
CGS [23]	VGG16	GAICD	0.773	54.7	72.0
Ours(basic)	VGG16	GAICD	0.744	52.0	70.5
Ours	VGG16	GAICD	**0.795**	**59.7**	**77.0**

we observe that $K = 9$ (row 2) achieves the best performance, because nine partitions can help capture more fine-grained partition-aware information. Besides, we evaluate some direct ways to leverage human bounding box for image cropping, yet producing poor results (see Supplementary).

Content-preserving Feature: Based on row 1, we add our content-preserving feature and report the results in row 7, in which we concatenate content-preserving feature with the region feature extracted from basic feature map. The results show the effectiveness of content-preserving feature. Next, we conduct ablation studies (row 8-11) based on row 7. We first replace GCN with conventional convolution layers (row 8) and observe the performance drop, which proves that it is useful to exploit the mutual relation between different regions. Then, we remove the predicted heatmap $\widetilde{\mathbf{H}}$ (row 9), resulting in significant performance drop, which highlights the importance of important content information. Additionally, we replace the proposed pseudo ground-truth heatmap with the saliency map detected by [13] in row 10 and obtain inferior performance. As discussed in Section 3.3, this can be attributed to that saliency may not accurately reflect the content of interest for human-centric images. We also try using the content-preserving feature alone. Specifically, we only use content-preserving feature $\mathbf{f}^c$ to predict the aesthetic score (row 11). The performance is even worse than row 1, because the content-preserving feature is lacking in detailed content information and thus insufficient for aesthetic prediction.

4.4 Comparison with the State-of-the-arts

Quantitative comparison: We compare the performance of our model with the state-of-the-art methods on 50 human-centric images of GAICD [43] dataset in

Table 3. Comparison with the state-of-the-art methods on human-centric images in FCDB [5] and FLMS [11] datasets. GAIC(ext) [44] is the extension of GAIC[43]. The results marked with * are obtained using the released models from original papers

Method	Backbone	Training Data	IoU↑	Disp↓
VFN* [6]	AlextNet	Flickr	0.5114	0.1257
VFN [6]	VGG16	CPC	0.6509	0.0876
VEN* [41]	VGG16	CPC	0.6194	0.0930
VEN [41]	VGG16	CPC	0.6670	0.0837
ASM-Net [37]	VGG16	CPC	0.7084	0.0755
LVRN* [28]	VGG16	CPC	0.7373	0.0674
GAIC(ext)* [44]	MobileNetV2	GAICD	0.7126	0.0724
GAIC(ext) [44]	VGG16	CPC	0.7260	0.0708
CGS [23]	VGG16	CPC	0.7331	0.0689
CACNet [15]	VGG16	FCDB	0.7364	0.0676
Ours(basic)	VGG16	CPC	0.7263	0.0695
Ours	VGG16	CPC	**0.7469**	**0.0648**

Table 2. For the baselines with released models, we evaluate their models on the test set and report the results (marked with *). However, their backbone and training data may be different from our setting.

For fair comparison, we use the pretrained VGG16 [34] as the backbone for all baselines and train them on GAICD dataset, based on their released code or our own implementation. For our method, we additionally report the results of a basic version ("Ours(basic)") without using partition-aware feature or content-preserving feature (row 1 in Table 1). It can be seen that Ours(basic) yields similar results with GAIC(ext) because they adopt the same region feature extractor (RoI+RoD). Among the baselines, GAIC(ext)* [44] and CGS [23] are two competitive ones, owning to the more advanced architecture and the exploitation of mutual relations between different crops. Finally, our model outperforms all the state-of-the-art methods, which demonstrates that our method is more well-tailored for the human-centric image cropping task.

Apart from GAICD dataset [43], we also collect 176 and 39 human-centric images from existing FCDB [5] and FLMS [11] datasets, respectively, and compare our method with the state-of-the-art methods on these two datasets in Table 3. Following [41,37], we train the model on CPC dataset [41], and use IoU and Disp as evaluation metrics. Additionally, we adopt the strategy in [28] to generate candidate crops and return the top-1 result as best crop without extra post-processing for all methods except CACNet [15], which is trained to regress the best crop directly. As shown in Table 3, our proposed model produces better results, but the performance gain is less significant than that on GAICD dataset. As claimed in [43], one possible reason is that the IoU based metrics used in FCDB and FLMS datasets are not very reliable for evaluating cropping

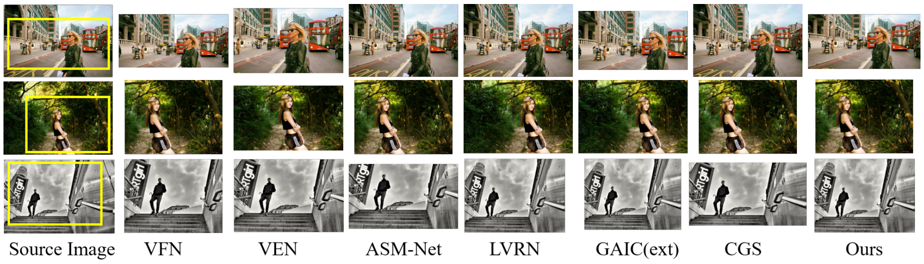

Fig. 4. Qualitative comparison of different methods on human-centric images. We show the best crops predicted by different methods, which demonstrate that our method can generate better results close to the ground-truth best crops (yellow)

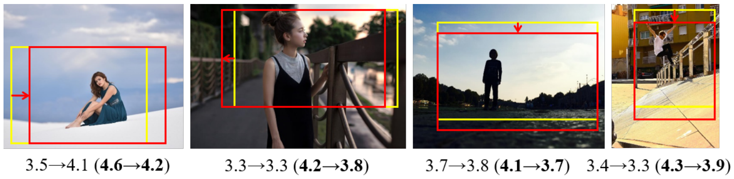

Fig. 5. Examples of partition-aware feature enhancing the discrimination power of basic feature. Given the ground-truth best crop of an image (yellow) and another crop with similar content (red), we show their scores predicted by using basic feature (out of bracket) and partition-aware feature (in bracket), respectively

performance. Furthermore, we also evaluate our method on both human-centric and non-human-centric images, and present results in Supplementary.

Qualitative comparison: We further conduct qualitative comparison to demonstrate the ability of our model in Figure 4. For each input image, we show the source image and the returned best crops by different methods, which demonstrates that our method can perform more reliable content preservation and removal. For example, in the first row of Figure 4, our method preserves more content on the left of human, probably because the person walks right to left, and reduces the top area that may hurt the image composition quality. In the second row, given the opposite face orientations to the first row, our model performs obviously different content preservation on the left/right sides of the human, yielding visually appealing crop. More qualitative results are shown in Supplementary.

4.5 Analysis of the Partition-aware Feature

To take a close look at the impact of partition-aware feature on candidate crop evaluation, we use the region features extracted from basic feature map and partition-aware feature map to predict scores for crops, respectively, corresponding to row 1 and row 2 in Table 1. As shown in Figure 5, to ensure that crop

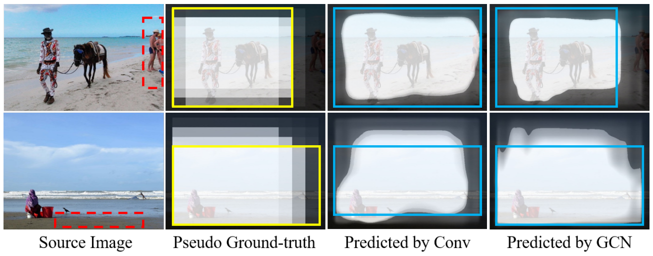

Fig. 6. Visualization of the heatmap that indicates aesthetically important content. We show the source image, its pseudo ground-truth heatmap, the heatmaps estimated by conventional convolution("Conv") / graph convolution ("GCN"). We also draw the ground-truth (*resp.*, predicted) best crops in yellow (*resp.*, blue) boxes

pairs have different aesthetic value yet similar content, for each image, we generate crop pair by moving its ground-truth best crop horizontally or vertically, in which the new crop still contains the human subject. We can see that using partition-aware feature consistently leads to larger and more reasonable score changes than basic feature despite the various face orientations or postures of the human in Figure 5, which is beneficial for correctly ranking crops with similar content.

4.6 Analysis of the Heatmap

The ablation study in Section 4.3 demonstrates the superiority of graph-based relation mining ("GCN") over the conventional convolution ("Conv") when predicting the heatmap of important content (see row 7,8 in Table 1). To reveal their difference qualitatively, we show the source image, its pseudo ground-truth heatmap, the heatmap predicted by "conv"/"graph" convolution in Figure 6. With GCN learning the mutual relation between different regions, the model can make a more reasonable estimation of important content, especially the border area. For example, in the source image in the first row, we show an unpleasant outer area in the red dashed box, which should be removed for composing a good crop. The unimportant content (low values in the heatmap) predicted by "GCN" completely covers the unpleasant area, while "Conv" only covers part of the unpleasant area. In the second row, unlike "Conv" that only deems the area around person as important, "GCN" predicts relatively high values for the area behind person, indicating that preserving such area in a crop may be beneficial. In summary, "GCN" can facilitate important content localization and contributes to more informative content-preserving feature.

5 User Study

Given the subjectiveness of aesthetic assessment task, we conduct user study to compare different methods, in which we employ total 265 human-centric images,

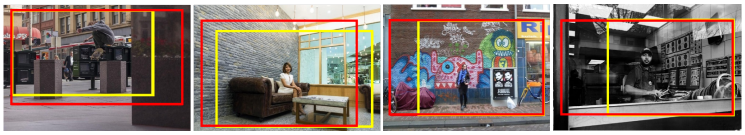

Fig. 7. Some failure cases in the test set of GAICD dataset [43]. For each image, the ground-truth and predicted best crops are drawn in yellow and red boxes, respectively

176 from FCDB [5], 50 from GAICD [43], and 39 from FLMS [11]. For each image, we generate 7 best crops by using seven different methods: VFN [6], VEN [41], ASM-Net [37], LVRN [28], GAIC(ext) [44], CGS [23], and our proposed method. 20 experts are invited to select the best result for each image. Then we calculate the percentage that the results generated by different methods are selected as the best ones. The percentages of the abovementioned six baselines are 1.7%, 5.6%, 9.0%, 14.6%, 15.7%, and 22.5%, respectively, while our method achieves the highest percentage 30.9% and clearly outperforms the other methods.

6 Limitations

Our method can generally produce reliable crops for human-centric images, but it still has some failure cases. Some failure cases in the test set of GAICD dataset [43] are shown in Figure 7, where the best crops produced by our method are far away from the ground-truth one and rank relatively low in the ground-truth annotations. For these examples, our method tends to preserve similar areas on the left/right side of the human subject in the best crop, probably because the complicated backgrounds and confusing human information (*e.g.*, inconsistent orientations between face and body) compromise the effectiveness of partition-aware feature and content-preserving feature.

7 Conclusion

In this paper, we have proposed a human-centric image cropping method with novel partition-aware and content-preserving features. The partition-aware feature allows to treat different regions in a candidate crop differently conditioned on the human information. The content-preserving feature represents how well each candidate crop preserves the important content. Extensive experiments have demonstrated that the proposed method can achieve the best performance on human-centric image cropping task.

Acknowledgement

The work is supported by Shanghai Municipal Science and Technology Key Project (Grant No. 20511100300), Shanghai Municipal Science and Technology Major Project, China (2021SHZDZX0102), and National Science Foundation of China (Grant No. 61902247).

References

1. Ardizzone, E., Bruno, A., Mazzola, G., et al.: Saliency based image cropping. In: ICIAP (2013)
2. Cavalcanti, C.S.V.C., Gomes, H., D, J.E.R., et al.: Combining multiple image features to guide automatic portrait cropping for rendering different aspect ratios. In: SITIS (2010)
3. Chen, H., Wang, B., Pan, T., Zhou, L., Zeng, H.: Cropnet: Real-time thumbnailing. In: ACMMM (2018)
4. Chen, J., Bai, G., Liang, S., Li, Z.: Automatic image cropping: A computational complexity study. In: CVPR (2016)
5. Chen, Y.L., Huang, T.W., Chang, K.H., Tsai, Y.C., Chen, H.T., Chen, B.Y.: Quantitative analysis of automatic image cropping algorithms: A dataset and comparative study. In: WACV (2017)
6. Chen, Y.L., Klopp, J., Sun, M., Chien, S.Y., Ma, K.L.: Learning to compose with professional photographs on the web. In: ACMMM (2017)
7. Chen, Z., Xu, Q., Cong, R., Huang, Q.: Global context-aware progressive aggregation network for salient object detection. In: AAAI (2020)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR (2009)
9. Deng, Y., Loy, C.C., Tang, X., et al.: Image aesthetic assessment: An experimental survey. IEEE Signal Processing Magazine **34**(4), 80–106 (2017)
10. Esmaeili, S.A., Singh, B., Davis, L.S.: Fast-at: Fast automatic thumbnail generation using deep neural networks. In: CVPR (2017)
11. Fang, C., Lin, Z., Mech, R., Shen, X.: Automatic image cropping using visual composition, boundary simplicity and content preservation models. In: ACMMM (2014)
12. Freeman, M.: The photographer's eye: Composition and design for better digital photos. CRC Press (2007)
13. Goferman, S., Zelnik-Manor, L., Tal, A.: Context-aware saliency detection. PAMI **10**(34), 1915–1926 (2012)
14. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV (2017)
15. Hong, C., Du, S., Xian, K., Lu, H., Cao, Z., Zhong, W.: Composing photos like a photographer. In: CVPR (2021)
16. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.H.: Deeply supervised salient object detection with short connections. In: CVPR (2017)
17. Hou, X., Zhang, L.: Saliency detection: A spectral residual approach. In: CVPR (2007)
18. Kao, Y., He, R., Huang, K., et al.: Automatic image cropping with aesthetic map and gradient energy map. In: ICASSP (2017)
19. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D., et al.: Visual Genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision **123**(1), 32–73 (2017)
20. Li, D., Wu, H., Zhang, J., Huang, K.: A2-RL: Aesthetics aware reinforcement learning for image cropping. In: CVPR (2018)
21. Li, D., Wu, H., Zhang, J., Huang, K.: Fast a3rl: Aesthetics-aware adversarial reinforcement learning for image cropping. TIP **28**(10), 5105–5120 (2019)
22. Li, D., Zhang, J., Huang, K.: Learning to learn cropping models for different aspect ratio requirements. In: CVPR (2020)

23. Li, D., Zhang, J., Huang, K., Yang, M.H.: Composing good shots by exploiting mutual relations. In: CVPR (2020)
24. Li, Q., Han, Z., Wu, X.M., et al.: Deeper insights into graph convolutional networks for semi-supervised learning. In: AAAI (2018)
25. Li, X., Li, X., Zhang, G., Zhang, X.: Image aesthetic assessment using a saliency symbiosis network. Journal of Electronic Imaging **28**(2), 023008 (2019)
26. Li, Z., Zhang, X.: Collaborative deep reinforcement learning for image cropping. In: ICME (2019)
27. Lu, P., Zhang, H., Peng, X., Jin, X.: An end-to-end neural network for image cropping by learning composition from aesthetic photos (2019)
28. Lu, W., Xing, X., Cai, B., Xu, X.: Listwise view ranking for image cropping. IEEE Access **7**, 91904–91911 (2019)
29. Mai, L., Jin, H., Liu, F., et al.: Composition-preserving deep photo aesthetics assessment. In: CVPR (2016)
30. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: NeurIPS (2019)
31. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. PAMI **39**, 1137–1149 (2015)
32. Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: The graph neural network model. IEEE Transactions on Neural Networks **20**(1), 61–80 (2008)
33. She, D., Lai, Y.K., Yi, G., Xu, K.: Hierarchical layout-aware graph convolutional network for unified aesthetics assessment. In: CVPR (2021)
34. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2015)
35. Su, Y.C., Vemulapalli, R., Weiss, B., Chu, C.T., Mansfield, P.A., Shapira, L., Pitts, C.: Camera view adjustment prediction for improving image composition. arXiv preprint arXiv:2104.07608 (2021)
36. Sun, J., Ling, H.: Scale and object aware image thumbnailing. IJCV **104**(2), 135–153 (2013)
37. Tu, Y., Niu, L., Zhao, W., Cheng, D., Zhang, L.: Image cropping with composition and saliency aware aesthetic score map. In: AAAI (2020)
38. Vig, E., Dorr, M., Cox, D.: Large-scale optimization of hierarchical features for saliency prediction in natural images. In: CVPR (2014)
39. Wang, W., Shen, J.: Deep cropping via attention box prediction and aesthetics assessment. In: ICCV (2017)
40. Wang, W., Shen, J., Ling, H.: A deep network solution for attention and aesthetics aware photo cropping. PAMI **41**(7), 1531–1544 (2018)
41. Wei, Z., Zhang, J., Shen, X., Lin, Z., Mech, R., Hoai, M., Samaras, D.: Good view hunting: Learning photo composition from dense view pairs. In: CVPR (2018)
42. Yan, J., Lin, S., Bing Kang, S., Tang, X.: Learning the change for automatic image cropping. In: CVPR (2013)
43. Zeng, H., Li, L., Cao, Z., Zhang, L.: Reliable and efficient image cropping: A grid anchor based approach. In: CVPR (2019)
44. Zeng, H., Li, L., Cao, Z., Zhang, L.: Grid anchor based image cropping: A new benchmark and an efficient model. PAMI **PP**(01), 1–1 (2020)
45. Zhang, L., Song, M., Yang, Y., Zhao, Q., Zhao, C., Sebe, N.: Weakly supervised photo cropping. TMM **16**(1), 94–107 (2013)
46. Zhang, L., Song, M., Zhao, Q., Liu, X., Bu, J., Chen, C.: Probabilistic graphlet transfer for photo cropping. TIP **22**(2), 802–815 (2012)

47. Zhang, M., Zhang, L., Sun, Y., Feng, L., Ma, W.: Auto cropping for digital photographs. In: ICME (2005)
48. Zhang, X., Li, Z., Constable, M., Chan, K.L., Tang, Z., Tang, G.: Pose-based composition improvement for portrait photographs. IEEE Transactions on Circuits and Systems for Video Technology **29**(3), 653–668 (2018)
49. Zhang, Y., Sun, X., Yao, H., Qin, L., Huang, Q.: Aesthetic composition represetation for portrait photographing recommendation. In: ICIP (2012)
50. Zhao, T., Wu, X.: Pyramid feature attention network for saliency detection. In: CVPR (2019)