Spatial-Semantic Collaborative Cropping for User Generated Content

Yukun Su1,2, Yiwen Cao1, Jingliang Deng1, Fengyun Rao2, Qingyao Wu1,3
1 School of Software and Engineering, South China University of Technology
2 WeChat, Tencent Inc.
3 Key Laboratory of Big Data and Intelligent Robot, Ministry of Education

Corresponding author.
suyukun666@outlook.com; fengyunrao@tencent.com; qyw@scut.edu.cn

Abstract
A large amount of User Generated Content (UGC) is uploaded to the Internet daily and displayed to people worldwide through the client side (e.g., mobile and PC). This requires the cropping algorithms to produce the aesthetic thumbnail within a specific aspect ratio on different devices. However, existing image cropping works mainly focus on landmark or landscape images, which fail to model the relations among the multi-objects with the complex background in UGC. Besides, previous methods merely consider the aesthetics of the cropped images while ignoring the content integrity, which is crucial for UGC cropping. In this paper, we propose a Spatial-Semantic Collaborative cropping network (S2CNet) for arbitrary user generated content accompanied by a new cropping benchmark. Specifically, we first mine the visual genes of the potential objects. Then, the suggested adaptive attention graph recasts this task as a procedure of information association over visual nodes. The underlying spatial and semantic relations are ultimately centralized to the crop candidate through differentiable message passing, which helps our network efficiently to preserve both the aesthetics and the content integrity. Extensive experiments on the proposed UGCrop5K and other public datasets demonstrate the superiority of our approach over state-of-the-art counterparts.

Introduction
Image cropping, with the aim to automatically excavate appealing views in photography, is widely used for image aesthetic compositions such as thumbnail generation (Chen et al. 2018; Esmaeili, Singh, and Davis 2017), shot recommendation (Li et al. 2018; Wei et al. 2018) and portrait suggestion (Zhang et al. 2018; Yee, Tantipongpipat, and Mishra 2021), etc. Among them, image thumbnailing or cover cropping is a vital application for the explosive emerging User Generated Content (UGC). Since users upload their self-created images or videos to the social media platform using different types of shooting equipment with lenses of various aspect ratios, as shown in Fig 1, this requires the cropping algorithms to generate the fixed aspect ratios cover images for content aesthetics and format unity.

However, several previous works (Chen et al. 2017b; Wei et al. 2018; Zeng et al. 2019, 2020; Pan et al. 2021; Jia et al. 2022) mainly focus on some iconic landscape images (Chen et al. 2017a; Zeng et al. 2020) collected from *Flickr* or even some human-centric (Zhang et al. 2022) images, where these images have clean backgrounds and they are relatively simple to crop. By contrast, the main challenges of cropping the user generated content are three folds: (i) UGC is more complex with different foreground multi-objects and chaotic backgrounds, thus it’s necessary to mine the relations between different objects to find the appealing crops. Meanwhile, some of the saliency-based cropping methods (Chen et al. 2016; Tu et al. 2020; Zhang et al. 2022; Cheng, Lin, and Allebach 2022) may fail to locate the accurate content; (ii) In addition to ensuring the aesthetics of the cropped images, content integrity is also crucial, which conveys the main message to the viewers. As shown in Fig 1(b), for some news clips or lyric videos, the cropping target should remain the main attributes of the image except for the people, such as the news headline and the complete lyrics. As for the multi-people images, incomplete face cases should be avoided; (iii) UGC cropping usually requires the fixed aspect ratio image output for display. Therefore, some anchor-generation-based methods (Hong et al. 2021; Jia et al. 2022) are unsuitable since they follow the process...
like object detection (Carion et al. 2020) and yield the crop candidates without aspect ratio constraint, which inevitably hinders their application in real-world scenarios.

The common practice of image cropping is to rank candidate views by data-driven methods using deep neural networks (Simonyan and Zisserman 2014; Sandler et al. 2018). Although numerous efforts have been made, due to the limitations of the public datasets and the technical solutions, many existing methods fail to achieve satisfactory performance for user generated content. To learn the relations between different image regions, an alternative solution is to utilize graph convolution networks (GCNs) (Chen et al. 2020). Li et al. (Li et al. 2020) exploited relations between different candidates with a graph-based module. However, it does not consider the instance-level cues in the images and the vanilla GCN may lead to the over-smoothing phenomenon. Pan et al. (Pan et al. 2021) adopted the vision transformer (ViT) (Dosovitskiy et al. 2021) to model the appearance and topological composition representations. Furthermore, the original ViT ignores the edge and spatial information between the unstructured data.

To tackle the issues mentioned above, we propose a Spatial-Semantic Collaborative cropping network ($S^2CNet$) to effectively crop arbitrary user generated content. Firstly, we mine the visual genes of the potential objects utilizing the off-the-shelf approach (Ren et al. 2015) to obtain region-of-interests (RoIs). Afterwards, they are used as the bias combined with the crop candidate and are fed into the proposed framework. Specifically, we design an adaptive attention graph where each RoI is viewed as the node and the correlation between each other is represented as the edge. Unlike prior works, we build the graph considering semantic and spatial collaborative information to capture both feature appearance and topological composition representations. Furthermore, we modify the graph convolution operation into a graph-aware attention module to efficiently model the high-order relations among each RoI, which recasts the network as a procedure of information association over visual nodes. The updated messages are ultimately centralized to the crop candidate for aesthetic score prediction. Furthermore, we also construct a large UGCrop5K dataset to fill the gap in the image cropping domain, which contains 450,000 exhaustive annotated candidate crops on 5,000 images varying in different topics (e.g., lecture, gaming, VR, and vlog, etc.). Massive experimental results on the UGCrop5K dataset and other public benchmarks all reveal the superiority and effectiveness of our proposed network, which can outperform the state-of-the-art methods while keeping a good trade-off between speed and accuracy. Our contribution can be summarized as follows:

- We experimentally investigate the limitations of the existing cropping algorithms and analyze the main challenges in real-life applications. We then construct a new UGCrop5K benchmark, to our best knowledge, which is the largest densely labeled cropping dataset with 450,000 high-quality annotated candidate crops.
- We propose an efficient $S^2CNet$ with a modified adaptive attention graph to capture the relations between different objects in the images. By exploiting both semantic and spatial information, we can produce aesthetic cropped images and maintain content integrity.
- Extensive experiments conducted on the proposed and other general datasets validate the merits of our approach against state-of-the-art cropping methods.

Related Work

Aesthetic Image Cropping. Most of the early conventional works (Suh et al. 2003; Stentiford 2007; Marchesotti, Cifarelli, and Curso 2009; Liu et al. 2010; Zhang et al. 2013; Fang et al. 2014) are based on hand-craft aesthetic features (Li et al. 2006; Ma and Guo 2004) and some criteria-based detection features such as face detection (Zhang et al. 2005) and eye tracking (Santella et al. 2006), etc. Later, benefiting from the deep learning models, more researchers pay attention to designing data-driven methods in various ways. VFN (Chen et al. 2017b) proposed an end-to-end deep ranking net to implicitly model images. Later, Wei et al. (Wei et al. 2018) constructed a comparative photo composition (CPC) dataset for pairwise learning. However, pairwise learning cannot provide sufficient evaluation metrics for image cropping as pointed in (Zeng et al. 2019). Recently, some works (Wang and Shen 2017; Li et al. 2019; Tu et al. 2020) exploited saliency detection (Goferman, Zelnik-Manor, and Tal 2011; Hou et al. 2017) to first locate the salient region and then generate candidate crops preserving the important content. However, some of the complex UGC images and landscape photos have multiple salient objects or even no salient ones, which may lead to cropping failure (Lu et al. 2019). Hong et al. (Hong et al. 2021) designed a dual branch network with the key composition map. However, it requires auxiliary composition datasets and manually defined rules, which lowers the upper bound of the usage. Zeng et al. (Zeng et al. 2019, 2020) introduced an efficient grid-based cropping method and proposed a densely annotated benchmark with new evaluation metrics. Jia et al. (Jia et al. 2022) formulated the task as object detection as DETR (Carion et al. 2020). However, this kind of anchor generation approach can not yield specific aspect ratio crops. More recently, HCIC (Zhang et al. 2022) proposed a specific content-aware human-centric approach, which hinders its application for general object photos.

Region-based Relations Mining. Region-based relations mining is popular in visual tasks, which is widely used in video classification (Wang and Gupta 2018), segmentation (Wang et al. 2019), tracking (Gao, Zhang, and Xu 2019) and image outpainting (Yang et al. 2022), etc. In the image cropping area, yet, rare works attempt to model the visual regional correlations. Although the most relevant approach CGS (Li et al. 2020) proposed to model the mutual relations between the candidates, the global feature of each crop ignores the instance-wise information inside or outside the candidate, which fails to explicitly compose the visual elements and decides what should be preserved or abandoned. Besides, all the aforementioned strategies are usually built on graph (Chen et al. 2020), where they merely consider the semantic message of each node while neglecting the spatial location information. Furthermore, the conventional graph
convolution networks will cause an over-smoothing problem when layers become deeper. TransView (Pan et al. 2021) later employed a vision transformer to capture image pixel-wise dependencies. However, not all tokens are equally important in vision tasks as stated in (Zeng et al. 2022). In addition, the self-attention mechanism in the transformer ignores the effective use of edge cues of nodes, such cues enjoy a good inductive bias (Goyal and Bengio 2022) for multi-object region learning. To alleviate these issues, we propose a modified adaptive attention graph to perform image cropping, which can well model the instance-level relations and find aesthetically pleasing crops.

Methodology

Network Overview

Our motivation is based on explicitly building the compositional relations among the crop candidate and all object proposals. According to this, the network learns what should be removed, partially reserved and reserved for an appealing crop that enjoys considerable content integrity. As shown in Fig 2(a), for the visual object outside the crop candidate (e.g., the old man on the far left), since the person looks to the left, making the content semantically irrelevant to the crop and its aesthetic contribution is thereby weak. For the elements inside the crop candidate, we attempt to make the network capture mutual visually significant dependencies. And for some uncertain background objects, we learn to preserve the attractive parts while removing the redundant parts. To achieve this goal, we adopt the adaptive attention graph (AAG) to model the scalable connections among the regional contents rather than using the plain transformer (Dosovitskiy et al. 2021; Pan et al. 2021) to represent the pair-wise relation among different nodes (e.g., \( x_i \) and \( x_j \)), and distribute different weights to the edges, we establish the relevance in an embedding space (Vaswani et al. 2017) to compute the feature appearance similarity matrix \( M_{a(i,j)} \) in \( \mathbb{R}^{(N+1) \times (N+1)} \) as:

\[
M_{a(i,j)} = \frac{\phi(x_i)^T \varphi(x_j)}{\sqrt{d}},
\]

where \( \phi(x) = W_\phi x + b_\phi \) and \( \varphi(x) = W_\varphi x + b_\varphi \) are two learnable linear functions that project the feature into the high-dimensional subspace.

Spatial Edges. In addition to building the semantic relations among the nodes, the spatial information should also be considered since it contains useful topological representation. Specifically, we view the center coordinate \( p_i = (p_i^x, p_i^y) \) of the node \( x_i \)'s bounding box as an initial spatial feature. To
this end, we explicitly model the spatial position connections of nodes in the following optional ways:

**DisDrop**: One assumption is that the nodes that are closer in space have more important information than nodes that are far away. And some of the distant nodes contribute less or even zero to aesthetic composition. In this way, we try to drop out the spatial relations of nodes whose distance exceeds a certain threshold and compute the spatial position matrix \( M_p \in \mathbb{R}^{(N+1)\times(N+1)} \) as follows:

\[
M_{p(i,j)} = \begin{cases} 
\psi(D(p_i, p_j)) & \text{if } D(p_i, p_j) \leq \epsilon \times wid \\
0 & \text{if } D(p_i, p_j) > \epsilon \times wid
\end{cases}
\]

where \( D(\cdot) \) is the Euclidean distance calculating function, \( \epsilon \) denotes the threshold, and \( wid \) is the width of image. \( \psi(\cdot) \) denotes a Multi-Layer Perceptron (MLP) layer that projects the 1-dimensional distance to a high-dimensional vector.

**DisEmb**: However, in some cases, the relations between the nodes that are farther away are also significant. For example, if two people are far away, but they are facing each other and have communication (e.g., greeting or having eye contact), then the relation between these two nodes is much stronger than the closer nodes but without any connection. Based on this observation, we formulate the spatial position matrix \( M_p \) as follows:

\[
M_{p(i,j)} = ||(W_{m1}p_i + b_{m}) - (W_{m2}p_j + b_{m})||_2^2,
\]

where \( W_{m1} \) and \( b_{m2} \) are the different learnable weight matrices and biases that embed the distance into a vector.

**Correlation Adjacency**. In order to jointly capture sufficient spatial-semantic information, we then construct a spatial-semantic correlation adjacency matrix \( A \in \mathbb{R}^{(N+1)\times(N+1)} \) combining both representations as follows:

\[
A_{(i,j)} = \frac{M_{a(i,j)} \cdot e^{M_{p(i,j)}}}{\sum_{j=1}^{N+1} M_{a(i,j)} \cdot e^{M_{p(i,j)}}},
\]

where normalization is performed for each element. Thus, we have \( A_{i,j} \sim [0, 1] \).

**Graph-Aware Attention Module**

After assembling the graph, we perform the feature extraction over the nodes. As aforementioned, we modify the standard GCN to the graph-aware attention operation similar to Transformer (Dosovitskiy et al. 2021) but merge the spatial-semantic features to generate the attention weights.

**Feature Aggregation Gate (FAG)**. As depicted in Fig 2(c), before calculating the self-attention of the node features, they are first fed into the feature aggregation gate to implicitly embed the information from the adjacency tensor. Specifically, we view the nodes as tokens. Considering the input feature \( X \) and the correlation adjacency tensor \( A \), the scheme of FAG is computed as follows:

\[
X = RELU(AZX),
\]

where \( Z \in \mathbb{R}^{(N+1)\times d} \) is the learnable weight matrix. The output features \( X \) aggregate the neighbouring node features, which can dynamically generate tokens with the appropriate importance to perform graph understanding.

**Spatial-Semantic Oriented Self-Attention (S^2O-SA)**. Afterwards, the outputs from FAG are viewed as queries \( Q \), and the original nodes are used as keys \( K \) and values \( V \). We then reformulate the self-attention as follows:

\[
S^2O-SA = softmax\left(\frac{KQ^T}{\sqrt{d}} + M_a + M_p\right)V.
\]

By injecting both spatial and semantic edge features, it endows the self-attention mechanism with semantic-aware and topology-aware structures that models the nodes non-equally. We omit the multi-head operation for clarity. In practice, we adopt several parallel multi-head attention to concatenate the features for better representation fusion. Generally, the whole process can be stacked into multi-layers as follows:

\[
X = FAG(X),
X' = S^2O-SA(LN(X)) + X,
X = FFN(LN(X')) + X',
\]

where \( LN(\cdot) \) indicates the LayerNorm (Ba, Kiros, and Hinton 2016) and \( FFN \) is the feed-forward network. Note that unlike the position encoding in the transformer that is added for sequential input data, the nodes in our paper are not arranged sequentially and are connected by edges. The proposed spatial edge encodes the structural information in the self-attention of a graph with the capability to modulate the distance-related receptive field.

**Network Optimization**

Ultimately, after obtaining the features from the adaptive attention graph, two layers of the MLPs are exploited to centralize the update message of all the nodes to the crop candidate to predict the aesthetic score. We first utilize the weighted smooth \( \ell_1 \) loss (Ren et al. 2015) for score regression as follows:

\[
\mathcal{L}_{pred} = \frac{1}{K} \sum_{i=1}^{K} \ell_1(y_i - \hat{y}_i),
\]

where \( K \) is the number of the crop candidate within an image, \( y_i \) and \( \hat{y}_i \) are the predicted and ground-truth score of the \( i \)-th candidate view, respectively.

In addition, following (Li et al. 2020; Zhang et al. 2022), we also use the ranking loss (Chen et al. 2017b) to explicitly learn the relative sorting orders between different crops as follows:

\[
\mathcal{L}_{rank} = \frac{\sum_{i,j} \max(0, \sigma(\hat{y}_i - \hat{y}_j) - (y_i - y_j))(\hat{y}_i - \hat{y}_j))}{K(K - 1)/2},
\]

where \( \sigma(\cdot) \) is the sign function. And the whole network is trained in an end-to-end manner.

**Experiment**

**Datasets and Metrics**

**Datasets**: To fill the gap in the image cropping domain of real-life applications, we construct a large dataset, term as
Figure 3: Statistics of the proposed UGCrop5K dataset, including (a) some visualization sample images, (b) taxonomic structure, (c) scatter plot of image width versus image height distribution with marker size indicating the number, and (d) histograms of the MOS.

<table>
<thead>
<tr>
<th>Spatial Edge</th>
<th>UGCrop5K</th>
<th>GAICv1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ACC5</td>
<td>ACC10</td>
</tr>
<tr>
<td>DisDrop $\epsilon = 0.1$</td>
<td>58.7</td>
<td>70.4</td>
</tr>
<tr>
<td>DisDrop $\epsilon = 0.2$</td>
<td>60.4</td>
<td>71.6</td>
</tr>
<tr>
<td>DisDrop $\epsilon = 0.3$</td>
<td>59.6</td>
<td>70.7</td>
</tr>
<tr>
<td>DisEmb</td>
<td>60.8</td>
<td>72.1</td>
</tr>
</tbody>
</table>

Table 1: Analysis of different spatial edges.

<table>
<thead>
<tr>
<th>Object Proposal</th>
<th>UGCrop5K</th>
<th>GAICv1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ACC5</td>
<td>ACC10</td>
</tr>
<tr>
<td>$N = 8$</td>
<td>59.9</td>
<td>71.2</td>
</tr>
<tr>
<td>$N = 10$</td>
<td><strong>60.8</strong></td>
<td><strong>72.1</strong></td>
</tr>
<tr>
<td>$N = 12$</td>
<td>60.6</td>
<td>71.9</td>
</tr>
<tr>
<td>$N = 15$</td>
<td>60.4</td>
<td>71.5</td>
</tr>
</tbody>
</table>

Table 2: Analysis of the object proposal number.

UGCrop5K. Specifically, we first collect parts of the user generated content from the opensourc databases, including KoNViD-1k (Hosu et al. 2017), LIVE-VQC (Sinno and Bovik 2018a,b; Z. Sinno and A.C. Bovik 2018), YouTube-UGC (Wang, Inguva, and Adsumilli 2019) and Bilibili (Ma et al. 2019) social video websites. Furthermore, we also provide approximately 500 self-made contents using different devices (e.g., shooting by iPhone 13 Pro Max, DJI Mavic 3 and Canon EOS R5 vertically or horizontally within various aspect ratios) from different scenarios to guarantee the variety of the proposed dataset. We then use HECATE (Song et al. 2016) to generate the top-3 cover images from the collected videos automatically. For data cleaning, we manually remove low-quality ambiguous images (i.e., pure colour images, highly similar content images, and blurry images) and reduce repetition. Particularly, we also remove images that are potentially not necessary for cropping. Finally, we have a total of 5,000 images with different aspect ratios covering different scenes, as shown in Fig 3(a) – (c).

Subsequently, 20 annotators are invited to our image composition annotation task, including 3 non-professional students, 10 medium-professional people engaged in art-related studies and 7 experienced workers in photography. We generate 90 predefined anchor boxes similar to (Zeng et al. 2020) for each image and develop an online website annotation tool instead of the annotation software (Zeng et al. 2020) that depends on the specific computer environment to ease the burden of the annotators. Concretely, annotators assign each predefined crop an integer score ranging from 1 to 5, with higher scores representing the better composition. Each crop needs to be rated by at least 5 people. We finally calculate the mean opinion score (MOS) for each candidate crop as its ground-truth quality score, and Fig 3(d) shows the histograms of the MOS. In general, we have 5,000 images with 450,000 high-quality annotated candidate crops in the dataset, and we split 4,200 images for training and 800 images for testing. Due to resource constraints, we conducted experiments on an early version of the dataset annotated by partial annotators. The complete dataset and results will be released in Github. To verify the generalization of our model, we also conduct experiments on other public image cropping benchmarks: GAICv1 (Zeng et al. 2019) and GAICv2 (Zeng et al. 2020) datasets.

**Metrics:** Following (Zeng et al. 2019, 2020), we adopt the averaged Spearman’s Rank-order Correlation Coefficient (SRCC) and the averaged top-$k$ accuracy ($ACC_k$) for both $k = 5$ and $k = 10$ as evaluation metrics instead of the unreliable Intersection-over-Union (IoU) metric.

**Implementation Details**

Following the existing methods (Zeng et al. 2020; Pan et al. 2021; Zhang et al. 2022), we adopt MobileNetV2 (Sandler et al. 2018) pretrained on ImageNet (Deng et al. 2009) as backbone to extract multi-scale feature map. The short side of the input sample is resized to 256 and maintains the aspect ratio. The aligned size of RoIAlign is set to $15 \times 15$ and the proposal number is set to 10 empirically. We stack 2 layers of the adaptive attention graphs with the multi-head number of 4. The network is optimized by AdamW with the learning rate of 1e-4 for 80 epochs. Data augmentations are similar to prior works (Zeng et al. 2020; Li et al. 2020), including random flipping, saturation, and lighting noise are adopted.
Ablation Analysis

Exploration of different spatial edges: As shown in Table 1, we first explore different constructions of spatial edge mentioned before. It shows that when \( \epsilon = 0.2 \), DisDrop can yield relatively good performance. Although different thresholds \( \epsilon \) can be altered, DisEmd can always outperform DisDrop. As aforementioned, the reason for the above observation is that the mutual contributions of visual features at different locations weakened by distance may miss some high-order spatial information. Therefore, in the following paper, we adopt the DisEmd strategy to build the spatial edge in a more holistic manner.

Exploration of object proposal number. We also explore the effect of different object proposal numbers, as shown in Table 2. When \( N = 10 \), we can achieve satisfactory results on both datasets. When \( N = 8 \), \( \text{ACC}_5 \) can be improved in GAICv1 benchmark since it does not contain many objects as in UGCrop5k. When continuously increasing the \( N \), the cropping performance will drop. We conclude that too many object proposal features are redundant, which may confuse the network for learning effective relations.

Exploration of different proposed components. Table 3 analyzes that each component can boost our network to varying degrees compared to the baseline (our baseline depends on pure Transformer block). Particularly, \( + \mathcal{M}_p \) obtains more improvement than \( + \mathcal{M}_a \), we conjecture that the topological spatial cues are more unique and useful in the self-attention operation. By combining all the components, we can achieve the best performance, which verifies the proposed modules are helpful and indispensable.

Exploration of different graphs. We further compare our proposed graph with the conventional GCN (Li et al. 2020) and GAT (Shaked Brody 2022). After replacing our graph with GCN, \( \text{ACC}_5 \) will drop from 60.8 to 56.2 and 58.6 on UGCrop5k and GAICv1 benchmarks, respectively. When replacing with GAT, \( \text{ACC}_5 \) will drop to 59.4 and 58.6 on two above benchmarks. This validates the effectiveness of our proposed adaptive attention graph.

Model interpretability: As shown in Fig 4, we show how our proposed adaptive attention graph encodes the information. The top-1 crop candidate and some detected object proposals are depicted in the leftmost image. By comparing the feature distribution maps (Van der Maaten and Hinton 2008), we can observe that the different regional features will diffuse or aggregate rather than in a mixed cluster. More specifically, the crop view features are closer to reserved features and have a certain degree of overlap with the partially reserved ones while far away from the removed features. This is because the relations between the crop view and aesthetically unnecessary contents are weakened via graph learning, and the edge weights between the crop view and potentially necessary contents are strengthened. This helps the network explicitly discriminate good and bad views through backpropagation learning with annotation scores.

Compare with the State-of-the-art Methods

Quantitative Results. As shown in Table 5, S2CNet can not only outperform state-of-the-art methods on the proposed challenging UGCrop5k dataset but also achieve satisfactory results on two other general GAICv1 and GAICv2 benchmarks. Nevertheless, our network can also achieve reliable results in general scenarios, which demonstrates the effectiveness of learning soundness of the proposed network. Note that SFRC (Wang et al. 2023) utilized additional unlabelled test data for training. For fair comparisons, we report its performance under the inductive setting. Besides, we also report the model complexity and runtime, all the experiments are executed on a single NVIDIA RTX 2080Ti GPU. Our method can process images at rates of 162.8 FPS while keeping competitive results, which guarantees the efficiency and practicality of the network.

Qualitative Analysis. Fig 5(a) shows the qualitative comparisons, from which we can observe that: (i) Our method
Table 5: Quantitative comparison to other state-of-the-art approaches on UGCrop5K, GAICv1 and GAICv2 datasets. The best performance is in bold, and the second-best is underlined. † indicates our re-implement results since the authors do not provide codes. Other results are derived from the open-source codes and the original papers.

Table 4 shows that our method can achieve the highest percentage and outperform the other methods.

Conclusion

In this paper, we introduce a spatial-semantic collaborative cropping network for user generated content and conduct a large densely labeled UGCrop5k dataset for follow-up research in the cropping domain. By exploring the semantic appearance and spatial topology information of different visual patches, we address the cropping task from a comprehensive perspective. Extensive experiments on different datasets show that our method outperforms the existing cropping approaches qualitatively and quantitatively.


can produce more aesthetically pleasing cropped views. They not only retain the main foreground of the photos but also can effectively preserve or remove some areas of the background to a greater extent for composition, and it is closer to the best annotated ground-truth; (ii) Our method can maintain image content integrity. As shown in the last row in Fig 5(a), although other methods successfully crop the main person and achieve a relatively good view, they lose some useful attributes of the image (i.e., A2-RL (Li et al. 2018), VEN/VPN (Wei et al. 2018) and GAICv2 (Zeng et al. 2020) cut out the important theme text of the news; HCIC (Zhang et al. 2022) even only keeps the main person), which may deliver incomplete information to readers.

Applications. In real-life applications, cropping is usually constrained. As shown in Fig 5(b), our model can find good views under different constraints, which demonstrates the ability of our model and meets the demand for UGC cropping, including cover image cropping, thumbnailing and icon generation, etc.

User Study. To evaluate the qualities of views within specific aspect ratios, we compare the proposed method with other approaches (e.g., Mars (Li, Zhang, and Huang 2020) and GAICv2 (Zeng et al. 2020)) that can also handle specific ratio cropping through the subjective user study. We randomly collect 100 images and 200 images from GAICv2 (Zeng et al. 2020) and UGCrop5K datasets. Then 15 volunteers are invited to select their favourite crop view from the results. Note that the experts are unaware of the views produced from which algorithms for fair comparisons. Table 4 shows that our method can achieve the highest percentage and outperform the other methods.
Acknowledgments

This work was supported by National Natural Science Foundation of China (NSFC) 62272172, Guangdong Basic and Applied Basic Research Foundation 2023A1515012920. This work is supported in part by a Tencent Research Grant and National Natural Science Foundation of China (No. 62176002).

References


Zeng, W.; Jin, S.; Liu, W.; Qian, C.; Luo, P.; Ouyang, W.; and Wang, X. 2022. Not all tokens are equal: Human-centric...


