Spatial-Semantic Collaborative Cropping for User Generated Content

Yukun Su[1,2], **Yiwen Cao**[1], **Jingliang Deng**[1], **Fengyun Rao**[2], **Qingyao Wu**[1,3]*

[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
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 worldwidely 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 (S^2CNet) 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 retain 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

(a): *UGC on mobile side* **(b):** *UGC on PC or Pad sides*

Figure 1: Illustrative example of cropping for UGC in a real-life application. Note that the original size is marked above each image. For intuitive explanation, the red dashed box indicates the cropped image produced by our algorithm for a fixed aspect ratio and the extraneous content is removed.

*Corresponding author.

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 visual element dependencies. Yet, 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*^{2}*CNet***) 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*^{2}*CNet*** 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 Csurka 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

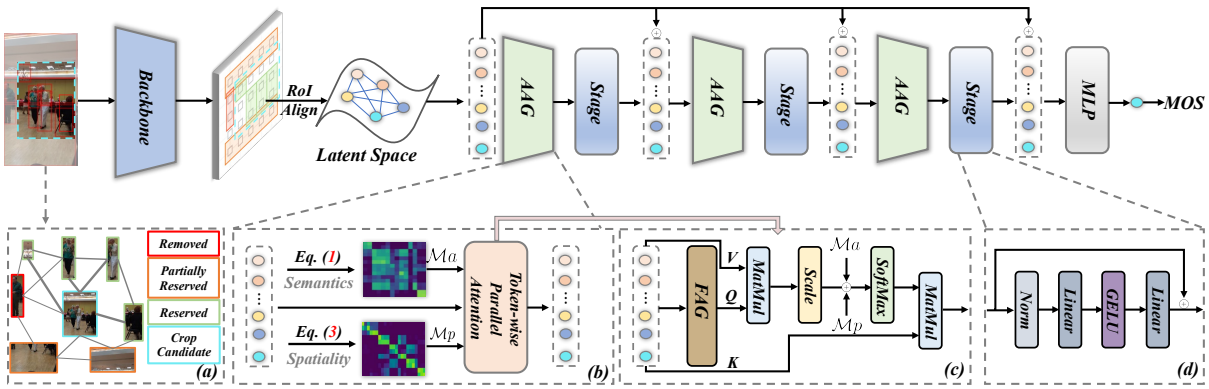

Figure 2: The overall pipeline of our proposed framework. We first use the convolutional backbone to extract visual features followed by RoIAlign (He et al. 2017) and RoDAlign (Zeng et al. 2019) extracting d-dimension features for each potential object and the crop candidate. These features are then provided as inputs to the proposed adaptive attention graph (AAG), which performs joint spatial-semantic information propagation over each node in the graph. Ultimately, the updated messages are centralized to the crop candidate node to perform aesthetic score prediction.

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 model the visual patches equally.

Concretely, given an input image $\mathcal{I}$ corresponding with the crop candidate, we leverage Faster RCNN (Ren et al. 2015) pretrained on Visual Genome (Krishna et al. 2017) to mine top-N potential visual objects. We then obtain the feature map $\mathcal{F}$ by passing the image into the convolutional backbone (Simonyan and Zisserman 2014; Sandler et al. 2018). After that, we apply RoIAlign (He et al. 2017) and RoDAlign (Gao, Zhang, and Xu 2019) operations followed by the FC layer to get d-dimensional features of the total visual regions as $X = [x_1, x_2, ..., x_{N+1}] \in \mathbb{R}^{(N+1)\times d}$ (N detection boxes and one crop candidate). These features are then fed into our proposed network to capture high-order information. Finally, we predict the aesthetic score by aggregating the updated features.

Adaptive Attention Graph

Formally, a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is defined as a set consisting of nodes and edges. Each node $v_i \in \mathcal{V}$ represents the extracted feature x_i, and each $e_{i,j} \in \mathcal{E}$ denotes the correlation between v_i and v_j. Note that we construct a fully connected graph since we consider that not only the relations between crop candidates and other regional objects are important, but the global relations among different visual objects also provide useful information for aesthetic composition.

Semantic Edges. 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 $\mathcal{M}_a \in \mathbb{R}^{(N+1)\times(N+1)}$ as:

$$\mathcal{M}_{a(i,j)} = \frac{\phi(x_i)^T \varphi(x_j)}{\sqrt{d}}, \tag{1}$$

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 $\mathcal{M}_p \in \mathbb{R}^{(N+1)\times(N+1)}$ as follows:

$$\mathcal{M}_{p(i,j)} = \begin{cases} \psi(\mathcal{D}(p_i, p_j)) & \text{if } \mathcal{D}(p_i, p_j) \leq \epsilon * wid \\ 0 & \text{if } \mathcal{D}(p_i, p_j) > \epsilon * wid \end{cases}, \quad (2)$$

where $\mathcal{D}(\cdot)$ is the *Euclidean* distance calculating function. ϵ 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 $\mathcal{M}_p$ as follows:

$$\mathcal{M}_{p(i,j)} = ||(W_m p_i + b_m) - (W_n p_j + b_n)||_2^2, \quad (3)$$

where $W_{m;n}$ and $b_{m;n}$ 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 $\mathcal{A} \in \mathbb{R}^{(N+1)\times(N+1)}$ combining both representations as follows:

$$\mathcal{A}_{(i,j)} = \frac{\mathcal{M}_{a(i,j)} \cdot e^{\mathcal{M}_{p(i,j)}}}{\sum_{j=1}^{N+1} \mathcal{M}_{a(i,j)} \cdot e^{\mathcal{M}_{p(i,j)}}}, \quad (4)$$

where normalization is performed for each element. Thus, we have $\mathcal{A}_{i,j} \sim [0, 1]$.

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 $\mathcal{A}$, the scheme of FAG is computed as follows:

$$X = RELU(\mathcal{A}ZX), \quad (5)$$

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:

$$S^2O\text{-}SA = \textit{softmax}(\frac{QK^T}{\sqrt{d}} + \mathcal{M}_a + \mathcal{M}_p)V. \quad (6)$$

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:

$$\begin{aligned} X &= FAG(X), \\ X' &= S^2O\text{-}SA(LN(X)) + X, \\ X &= FFN(LN(X')) + X', \end{aligned} \quad (7)$$

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 ℓ_1 loss (Ren et al. 2015) for score regression as follows:

$$\mathcal{L}_{pred} = \frac{1}{K}\sum_{i=1}^{K} \ell_1(y_i - \hat{y}_i), \quad (8)$$

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:

$$\mathcal{L}_{rank} = \frac{\sum_{i,j} \max(0, \sigma(\hat{y}_i - \hat{y}_j)((y_i - y_j)(\hat{y}_i - \hat{y}_j)))}{K(K-1)/2}, \quad (10)$$

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

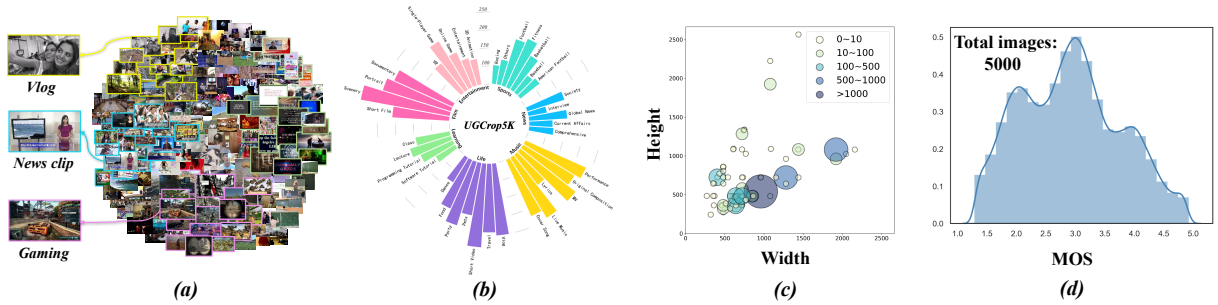

Figure 3: Statistics of the proposed *UGCrop5K* dataset, including (a) some visualization sample images, (b) taxonomic structure, (c) scatter plot of image width versus image height distribution with marker size indicating the number, and (d) histograms of the MOS.

Spatial Edge		***UGCrop5K***		***GAIC*v1**	
		ACC_5	ACC_{10}	ACC_5	ACC_{10}
DisDrop	$\epsilon = 0.1$	58.7	70.4	59.6	77.2
	$\epsilon = 0.2$	60.4	71.6	60.5	77.8
	$\epsilon = 0.3$	59.6	70.7	59.9	77.4
DisEmb		**60.8**	**72.1**	**61.0**	**78.1**

Table 1: Analysis of different spatial edges.

Object Proposal	***UGCrop5K***		***GAIC*v1**	
	ACC_5	ACC_{10}	ACC_5	ACC_{10}
$N = 8$	59.9	71.2	**61.2**	78.0
$N = 10$	**60.8**	**72.1**	61.0	**78.1**
$N = 12$	60.6	71.9	59.6	77.4
$N = 15$	60.4	71.5	58.7	77.0

Table 2: Analysis of the object proposal number.

UGCrop5K. Specifically, we first collect parts of the user generated content from the opensource 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) $\sim$ (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: ***GAIC*v1** (Zeng et al. 2019) and ***GAIC*v2** (Zeng et al. 2020) datasets.

***Metrics*:** Following (Zeng et al. 2019, 2020), we adopt the averaged Spearman's Rank-order Correlation Coefficient ($\overline{SRCC}$) and the averaged top-k accuracy ($\overline{ACC_k}$) 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 15×15 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.

$\mathcal{G}$	$\mathcal{M}_a$	$\mathcal{M}_p$	*UGCrop5K*		*GAICv1*	
			$\overline{ACC_5}$	ACC_{10}	$\overline{ACC_5}$	ACC_{10}
			54.3	64.8	49.7	68.4
✔			55.2	65.7	51.1	70.5
	✔		54.7	65.3	50.8	70.2
		✔	54.9	65.8	53.4	72.6
✔	✔		58.7	69.6	56.7	76.1
✔		✔	59.2	70.9	58.9	77.0
	✔	✔	60.0	71.5	60.2	77.4
✔	✔	✔	**60.8**	**72.1**	**61.0**	**78.1**

Table 3: Analysis of different proposed components. $\mathcal{G}$ indicates FAG module.

Ratios	*Perc. (%)*		
	Ours	Mars	GAICv2
3:4	**47.7**	35.3	17.0
4:3	**44.0**	24.3	31.7
16:9	**52.0**	23.3	24.7

Table 4: User study results.

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 ϵ 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 $N = 8$, $\overline{ACC_5}$ can be improved in *GAIC*v1 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, $\overline{ACC_5}$ will drop from **60.8** $\xrightarrow{-4.6}$ **56.2** and **61.0** $\xrightarrow{-3.7}$ **57.3** on *UGCrop5k* and *GAIC*v1 benchmarks, respectively. When replacing with GAT, $\overline{ACC_5}$ will drop to **59.4** and **58.6** on two above benchmarks. This validates the effectiveness of our proposed adaptive attention graph.

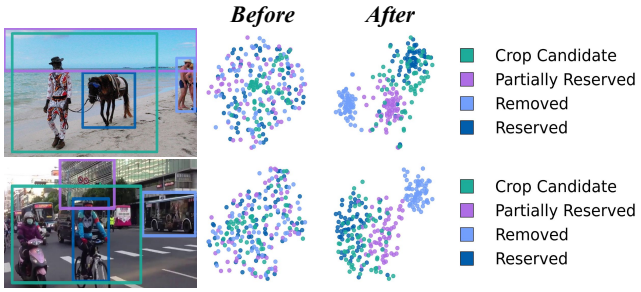

Figure 4: The t-SNE feature visualization before and after the proposed graph. Different colours indicate the crop candidates, the regions should be removed, reserved, or partially reserved, respectively. The features before the graph show indistinguishable clusters, while the features learned by our graph are more discriminative, which can guide the model to find good views more reasonably. Zoom in for the best view.

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 can help 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, ***S*^{2}*CNet*** can not only outperform state-of-the-art methods on the proposed challenging *UGCrop5k* dataset but also achieve satisfactory results on two other general *GAIC*v1 and *GAIC*v2 benchmarks. Nevertheless, our network can also achieve reliable results in general scenarios, which demonstrates the effectiveness and soundness of the proposed network. Note that SFRC (Wang et al. 2023) utilized additional unlabeled 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

Model	*Param (M)*	*FPS* ↑	***UGCrop5K***			***GAICv1***			***GAICv2***		
			$SRCC$	ACC_5	ACC_{10}	$SRCC$	ACC_5	ACC_{10}	$SRCC$	ACC_5	ACC_{10}
VFN	11.55	0.4	0.372	25.4	36.1	0.450	26.7	38.7	0.485	26.4	40.1
A2-RL	24.11	2.6	-	22.8	33.9	-	23.0	38.5	-	23.2	39.5
VEN	40.93	0.3	0.394	31.3	42.7	0.621	37.6	50.9	0.616	35.5	48.6
VPN	65.31	96.2	-	35.8	44.3	-	40.0	49.5	-	36.0	48.5
GAICv1	13.54	129.8	0.418	45.7	52.8	0.735	46.6	65.5	0.832	63.5	79.0
GAICv2	**1.81**	**212.4**	0.466	54.7	64.5	0.783	57.2	75.5	0.849	63.9	79.7
ASM-Net	14.95	102.0	0.435†	52.8†	63.2†	0.766	54.3	71.5	0.837†	63.2†	79.1†
CGS	13.68	100.0	0.467	56.4	66.8	**0.795**	59.7	77.8	0.848	63.5	79.4
TransView	4.62	147.3	0.482†	57.9†	69.4†	0.789†	59.2†	77.4†	0.857	63.9	82.4
HCIC	19.47	128	0.449	54.5	64.1	0.793	58.6	74.5	0.851	63.8	81.3
SFRC	5.91	40	-	-	-	-	-	-	**0.865**	63.7	82.6
S²CNet	3.92	162.8	**0.502**	**60.8**	**72.1**	0.793	**61.0**	**78.1**	0.861	**64.0**	**82.7**

Table 5: Quantitative comparison to other state-of-the-art approaches on *UGCrop5K*, *GAIC*v1 and *GAIC*v2 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.

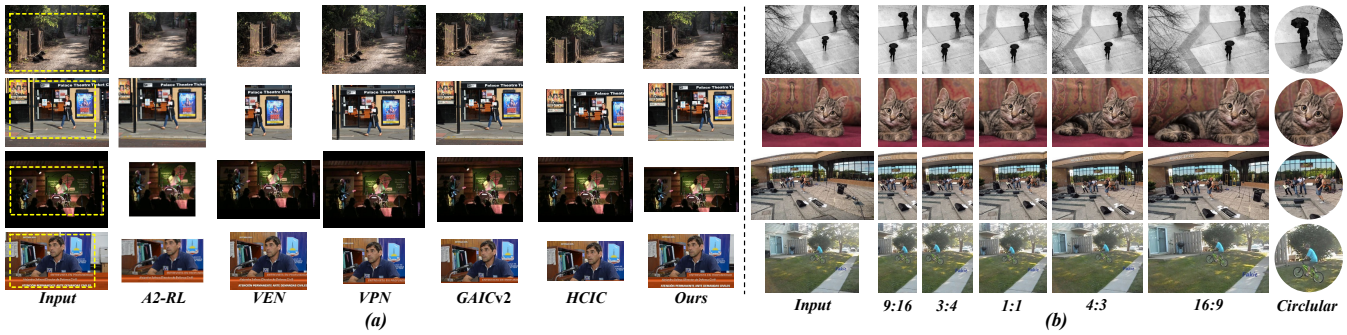

Figure 5: (a): Qualitative comparisons of different state-of-the-art methods. The first two rows of images are from the *GAIC*v1 and *GAIC*v2 datasets, and the last two rows of images are from the *UGCrop5k* dataset. The top-scored best crops are in the yellow dotted box. (b): Image cropping results with different aspect ratios.

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 *GAIC*v2 (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.

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.

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

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, 213–229. Springer.

Chen, H.; Wang, B.; Pan, T.; Zhou, L.; and Zeng, H. 2018. CropNet: Real-time thumbnailing. In *Proceedings of the 26th ACM international conference on Multimedia*, 81–89.

Chen, J.; Bai, G.; Liang, S.; and Li, Z. 2016. Automatic image cropping: A computational complexity study. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 507–515.

Chen, M.; Wei, Z.; Huang, Z.; Ding, B.; and Li, Y. 2020. Simple and deep graph convolutional networks. In *International conference on machine learning*, 1725–1735. PMLR.

Chen, Y.-L.; Huang, T.-W.; Chang, K.-H.; Tsai, Y.-C.; Chen, H.-T.; and Chen, B.-Y. 2017a. Quantitative Analysis of Automatic Image Cropping Algorithms:A Dataset and Comparative Study. In *IEEE WACV 2017*.

Chen, Y.-L.; Klopp, J.; Sun, M.; Chien, S.-Y.; and Ma, K.-L. 2017b. Learning to compose with professional photographs on the web. In *Proceedings of the 25th ACM international conference on Multimedia*, 37–45.

Cheng, Y.; Lin, Q.; and Allebach, J. P. 2022. Re-Compose the Image by Evaluating the Crop on More Than Just a Score. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1–9.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR*.

Esmaeili, S. A.; Singh, B.; and Davis, L. S. 2017. Fast-at: Fast automatic thumbnail generation using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4622–4630.

Fang, C.; Lin, Z.; Mech, R.; and Shen, X. 2014. Automatic image cropping using visual composition, boundary simplicity and content preservation models. In *Proceedings of the 22nd ACM international conference on Multimedia*, 1105–1108.

Gao, J.; Zhang, T.; and Xu, C. 2019. Graph convolutional tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4649–4659.

Goferman, S.; Zelnik-Manor, L.; and Tal, A. 2011. Context-aware saliency detection. *IEEE transactions on pattern analysis and machine intelligence*, 34(10): 1915–1926.

Goyal, A.; and Bengio, Y. 2022. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, 478(2266): 20210068.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Hong, C.; Du, S.; Xian, K.; Lu, H.; Cao, Z.; and Zhong, W. 2021. Composing photos like a photographer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7057–7066.

Hosu, V.; Hahn, F.; Jenadeleh, M.; Lin, H.; Men, H.; Szirányi, T.; Li, S.; and Saupe, D. 2017. The Konstanz natural video database (KoNViD-1k). In *2017 Ninth international conference on quality of multimedia experience (QoMEX)*, 1–6. IEEE.

Hou, Q.; Cheng, M.-M.; Hu, X.; Borji, A.; Tu, Z.; and Torr, P. H. 2017. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3203–3212.

Jia, G.; Huang, H.; Fu, C.; and He, R. 2022. Rethinking Image Cropping: Exploring Diverse Compositions from Global Views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2446–2455.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123: 32–73.

Li, D.; Wu, H.; Zhang, J.; and Huang, K. 2018. A2-RL: Aesthetics aware reinforcement learning for image cropping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8193–8201.

Li, D.; Zhang, J.; and Huang, K. 2020. Learning to learn cropping models for different aspect ratio requirements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12685–12694.

Li, D.; Zhang, J.; Huang, K.; and Yang, M.-H. 2020. Composing good shots by exploiting mutual relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4213–4222.

Li, J.; Datta, R.; Joshi, D.; and Wang, J. 2006. Studying aesthetics in photographic images using a computational approach. *Lecture Notes in Computer Science*, 3953: 288–301.

Li, X.; Li, X.; Zhang, G.; and Zhang, X. 2019. Image aesthetic assessment using a saliency symbiosis network. *Journal of Electronic Imaging*, 28(2): 023008–023008.

Liu, L.; Chen, R.; Wolf, L.; and Cohen-Or, D. 2010. Optimizing photo composition. In *Computer graphics forum*, volume 29, 469–478. Wiley Online Library.

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

Ma, M.; and Guo, J. K. 2004. Automatic image cropping for mobile device with built-in camera. In *First IEEE Consumer Communications and Networking Conference, 2004. CCNC 2004.*, 710–711. IEEE.

Ma, S.; Cui, L.; Dai, D.; Wei, F.; and Sun, X. 2019. LiveBot: Generating Live Video Comments Based on Visual and Textual Contexts. In *AAAI 2019*.

Marchesotti, L.; Cifarelli, C.; and Csurka, G. 2009. A framework for visual saliency detection with applications to image thumbnailing. In *2009 IEEE 12th International Conference on Computer Vision*, 2232–2239. IEEE.

Pan, Z.; Cao, Z.; Wang, K.; Lu, H.; and Zhong, W. 2021. Transview: Inside, outside, and across the cropping view boundaries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4218–4227.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4510–4520.

Santella, A.; Agrawala, M.; DeCarlo, D.; Salesin, D.; and Cohen, M. 2006. Gaze-based interaction for semi-automatic photo cropping. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, 771–780.

Shaked Brody, E. Y., Uri Alon. 2022. How Attentive are Graph Attention Networks? In *International Conference on Learning Representations*.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sinno, Z.; and Bovik, A. C. 2018a. Large-scale study of perceptual video quality. *IEEE Transactions on Image Processing*, 28(2): 612–627.

Sinno, Z.; and Bovik, A. C. 2018b. Large scale subjective video quality study. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, 276–280. IEEE.

Song, Y.; Redi, M.; Vallmitjana, J.; and Jaimes, A. 2016. To click or not to click: Automatic selection of beautiful thumbnails from videos. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, 659–668.

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

Suh, B.; Ling, H.; Bederson, B. B.; and Jacobs, D. W. 2003. Automatic thumbnail cropping and its effectiveness. In *Proceedings of the 16th annual ACM symposium on User interface software and technology*, 95–104.

Tu, Y.; Niu, L.; Zhao, W.; Cheng, D.; and Zhang, L. 2020. Image cropping with composition and saliency aware aesthetic score map. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 12104–12111.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, C.; Niu, L.; Zhang, B.; and Zhang, L. 2023. Image Cropping With Spatial-Aware Feature and Rank Consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10052–10061.

Wang, W.; Lu, X.; Shen, J.; Crandall, D. J.; and Shao, L. 2019. Zero-shot video object segmentation via attentive graph neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9236–9245.

Wang, W.; and Shen, J. 2017. Deep cropping via attention box prediction and aesthetics assessment. In *Proceedings of the IEEE International Conference on Computer Vision*, 2186–2194.

Wang, X.; and Gupta, A. 2018. Videos as space-time region graphs. In *Proceedings of the European conference on computer vision (ECCV)*, 399–417.

Wang, Y.; Inguva, S.; and Adsumilli, B. 2019. YouTube UGC dataset for video compression research. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, 1–5. IEEE.

Wei, Z.; Zhang, J.; Shen, X.; Lin, Z.; Mech, R.; Hoai, M.; and Samaras, D. 2018. Good view hunting: Learning photo composition from dense view pairs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5437–5446.

Yang, C.-A.; Tan, C.-Y.; Fan, W.-C.; Yang, C.-F.; Wu, M.-L.; and Wang, Y.-C. F. 2022. Scene graph expansion for semantics-guided image outpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15617–15626.

Yee, K.; Tantipongpipat, U.; and Mishra, S. 2021. Image cropping on twitter: fairness metrics, their limitations, and the importance of representation, design, and agency. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–24.

Z. Sinno and A.C. Bovik. 2018. https://live.ece.utexas.edu/research/LIVEVQC/index.html.

Zeng, H.; Li, L.; Cao, Z.; and Zhang, L. 2019. Reliable and efficient image cropping: A grid anchor based approach. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5949–5957.

Zeng, H.; Li, L.; Cao, Z.; and Zhang, L. 2020. Grid anchor based image cropping: A new benchmark and an efficient model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3): 1304–1319.

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

visual analysis via token clustering transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11101–11111.

Zhang, B.; Niu, L.; Zhao, X.; and Zhang, L. 2022. Human-Centric Image Cropping with Partition-Aware and Content-Preserving Features. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, 181–197. Springer.

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

Zhang, M.; Zhang, L.; Sun, Y.; Feng, L.; and Ma, W. 2005. Auto cropping for digital photographs. In *2005 IEEE international conference on multimedia and expo*, 4–pp. IEEE.

Zhang, X.; Li, Z.; Constable, M.; Chan, K. L.; Tang, Z.; and Tang, G. 2018. Pose-based composition improvement for portrait photographs. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(3): 653–668.