TransView: Inside, Outside, and Across the Cropping View Boundaries

Zhiyu Pan¹  Zhiguo Cao¹  Kewei Wang¹  Hao Lu¹,*  Weicai Zhong²

¹Key Laboratory of Image Processing and Intelligent Control, Ministry of Education
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
²Huawei CBG Consumer Cloud Service Search Product & Big Data Platform Department
{zhiyupan, hlu}@hust.edu.cn

Abstract

We show that relation modeling between visual elements matters in cropping view recommendation. Cropping view recommendation addresses the problem of image recomposition conditioned on the composition quality and the ranking of views (cropped sub-regions). This task is challenging because the visual difference is subtle when a visual element is reserved or removed. Existing methods represent visual elements by extracting region-based convolutional features inside and outside the cropping view boundaries, without probing a fundamental question: why some visual elements are of interest or of discard? In this work, we observe that the relation between different visual elements significantly affects their relative positions to the desired cropping view, and such relation can be characterized by the attraction inside/outside the cropping view boundaries and the repulsion across the boundaries. By instantiating a transformer-based solution that represents visual elements as visual words and that models the dependencies between visual words, we report not only state-of-the-art performance on public benchmarks, but also interesting visualizations that depict the attraction and repulsion between visual elements, which may shed light on what makes for effective cropping view recommendation.

1. Introduction

Image composition is one of key factors in professional photography. The term ‘composition’ can be considered ‘the organization of the elements of art’ [49]. Sad to say, the skills and tricks for organizing visual elements are the main barrier that prevents ordinary people from taking professional photos. Nonetheless, many amateurs are still eager to compose photos like a photographer, even without expertise and training. The demand for automatic composition has thus come into the eye of computer vision community, and much effort has been made to solve image recomposition [8, 18, 27, 38, 48, 51].

One of off-the-shelf and low-end technologies for image recomposition [20] is image cropping. It aims to find the most aesthetic view (a sub-region defined by the cropping box) in an image. The typical paradigm of image cropping is to rank candidate views and to retrieve appropriate ones. This task is also given the name of cropping view recommendation. A straightforward idea is to score and rank candidate views by artificially designed evaluation criteria. However, such criteria cannot cover the principles of art and align poorly with the actual preference of users. Recently another promising way is to learn directly from data [31, 48, 51]. In particular, convolutional models are developed as possible solutions to the dilemma above. These data-driven methods predict scores conditioned on region-aware features that delineate the presence of visual elements (Fig. 1(a)). In this way, these methods can be viewed as finding connections between the presence of vi-
sual elements and the good composition. However, according to the definition of composition, we argue that what explains why a cropping view is of good composition is not the presence of visual elements, but the harmony in the organization between them. Since the organization is often interpreted as relation between elements [12], composition patterns should be found in the dependencies between visual elements.

Convolutional networks, however, are weak in modeling dependencies. First, long-range dependencies can only be encoded when the receptive field is sufficiently large. Second, the difficulty in optimization [17, 36] causes multi-hop dependency modeling [46] such that messages are hard to travel between distant positions. This is also why the empirical receptive field is limited [29]. In the field of natural language processing [1, 39, 42], this problem has been studied in depth and addressed well with the transformer architecture [42]. Transformer [42] can precisely model all pairwise dependencies in a parallel manner. Since modeling parallel relation is important in cropping view recommendation, we believe transformer can be an effective tool to mine relation-aware composition patterns.

In this work, we propose to explicitly encode the dependencies of visual elements inside, outside, and across the cropping view boundaries (Fig. 1(b)). In particular, we borrow the concept of ‘visual words’ in [11] to represent visual elements and model pairwise dependencies between the visual words via repeated attentional operators [1, 22]. We intend to characterize two forms of dependencies: the attraction dependency and the repulsion dependency. The attraction dependency aims to contribute to the global harmony between expected foreground visual words, e.g., the two persons in Fig. 1(b), or between aesthetically necessary background visual words, e.g., the surrounding ice and glacier; the repulsion dependency is used to depict the semantically/spatially incompatible relation against the desired visual words, e.g., the superfluous ice land that weakens the role of the two persons. We believe the two dependencies can be used as a criterion to judge a candidate view: a desirable cropping box should not only reserve the main elements clustered by attraction but also discard the elements repulsive to the main subject. To implement the criterion, we propose the TransView model that encodes three types of dependencies: attraction inside the cropping boundaries, attraction outside the boundaries, and repulsion across the boundaries.

Experimental results on the public benchmarks show that TransView outperforms state-of-the-art region-feature-based methods. We also show that TransView does model the attraction and repulsion without supervision via interpretable activation maps. Further feature visualizations show that explicit encoding of the attraction and repulsion leads to distinguishable features of similar views.

2. Related Work

We review cropping-based image recomposition and attention-based relation encoding.

Cropping Based Image Recomposition. How to automatically recompose the image by cropping is driven by two main ideas: artificial-criteria-driven cropping and data-driven cropping. Targeting image thumbnail [4, 30, 38] and aesthetic cutting [8, 15], criteria-based cropping methods extract features by detection (e.g., saliency detection [43], face detection [37], and text detection [5]), by eye fixation data [35], or by predefined composition rules [8, 13, 32, 50, 53] such as the rule of thirds. By integrating hand-craft features into an energy function, views can be evaluated by function scores. These methods, however, can only generate cropping boxes with dominant subjects. Recently, data-driven cropping models emerge. These models generally follow a two-stage pipeline. First, candidate views are generated following aesthetic prior knowledge [52]. Then, candidate views are ranked according to learned expert knowledge, which is often modeled by self-supervision [7], saliency prediction [21, 40] followed by aesthetic evaluation [44, 45], knowledge distillation [48], RoI and RoD feature fusion [51, 52], and mutual relation mining [27]. Some other works generate cropping boxes directly by reinforcement learning [24, 25] and meta-learning [26]. One common deficiency of the methods above is that they only consider the content of regions rather than answering the question why some visual elements are of interest or of discard. Our work, by contrast, recomposes images by modeling relation between visual elements.

Relation Encoding by Attention. Encoding relation between elements in sequences has been widely studied in the field of natural language processing [1, 39, 42]. Recurrent neural networks [9, 36] factor dependencies along the input direction of signals, which precludes the parallel nature of some sequential elements [42]. The attention based transformer [42], however, can model global dependencies parallelly. The superiority of transformer makes it suitable for long sequences and has shown applications in computer vision. ViT [11] is a classic transformer encoder that serializes image patches and reports state-of-the-art performance on image classification. DETR [2] and its variant [54] transform object detection into a set prediction problem and exploit the relation between patch features and object queries. Hand transformer [19] tackles the difficulty of modeling structural dependencies in 3D hand pose estimation. Moreover, many dense prediction problems [28], such as semantic segmentation [47], image restoration [3], and image generation [33], also benefit from the transformer architecture. Inspired by the idea of image-sequence transduction, we propose the transformer-based model for processing visual elements informed by attractive and repulsive visual words.
3. Attraction-Repulsion-Aware Recomposition

3.1. Overview

According to the definition of composition in Section 1, the art of composition is about the harmony in the organization of visual elements. An alternative interpretation is that the composition quality of one image can be assessed by the organization of visual elements. Inspired by this, we propose to model the organization of visual elements explicitly to evaluate the composition quality. By delineating the organization as dependencies between visual elements, we present TransView, an attraction-repulsion-aware model based on the transformer architecture. TransView includes three main components: a convolutional backbone, two attraction learners, and a repulsion learner. The technical pipeline is illustrated in Fig. 2.

The convolution backbone generates the visual words. Following existing practices [51, 52] that consider not only the region of interest (RoI) but also the region of discard (RoD) [51], visual words are divided into two types: internal and external words. As aforementioned, for a desirable cropping box, visual words in the same region should be attractive, and words from different regions are expected to be repulsive. The transformer encoder, which inputs a single sequence, is suitable to encode the attraction between internal/external words; and the transformer decoder, which requires two sequences input, is appropriate to encode the repulsion between internal and external words. Hence, the transformer encoder and decoder are adopted as the attraction learners and the repulsion learner, respectively. By concatenating the attraction-repulsion-aware features with the region-aware one, the final assessment for a view can be obtained.

3.2. Representing Images by Visual Words

Digital images are represented by pixels. Despite pixels can be sequentialized, the curse of dimensionality and the uncertainty of sequence length make them difficult to deal with. Following [2], we represent visual words using a MobileNetv2-based backbone [34] with RoIAlign [16] and RoDAlign [51] operators.

By downsampling the feature map, each local region of the feature map corresponds to an image patch. We follow [51] to extract the multi-scale feature $F$ before RoIAlign and RoDAlign. After RoIAlign and RoDAlign, the RoI and RoD features can be extracted and aligned to $F_{RoI} \in \mathbb{R}^{D \times H \times W}$ and $F_{RoD} \in \mathbb{R}^{D \times H \times W}$, respectively. By flattening $F_{RoI}$ and $F_{RoD}$ into sequences, we can obtain fixed-length internal words $C^I \in \mathbb{R}^{D \times N}$ and external words $C^E \in \mathbb{R}^{D \times N}$, where $N = H \times W$.

However, representing images by visual words per se is not sufficient. We also follow [14, 42] to add positional embeddings to the visual words to supplement spatial information. In particular, learnable internal positional embedding $P^I \in \mathbb{R}^{D \times N}$ and external positional embedding $P^E \in \mathbb{R}^{D \times N}$ are used to learn different spatial structures of internal and external visual words. In this way, we obtain the content- and position-aware internal visual words $X^I \in \mathbb{R}^{D \times N}$ and external visual words $X^E \in \mathbb{R}^{D \times N}$ by $X^I = C^I + P^I$ and $X^E = C^E + P^E$, respectively.

3.3. Attraction Modeling

Most principles of art describe good composition as the patterns of dependencies between visual elements [12]. According to this definition, we factorize the dependencies into the patterns of attraction between internal/external visual words and the patterns of repulsion between internal and external words. For the attraction dependency, the transformer...
encoder is employed as the attraction learner. The input is a sequence $X_0^I$ formed by internal visual words, defined by

$$X_0^I = \{x_{(0)}^I, x_{(1)}^I, \ldots, x_{(N)}^I\}, \quad (1)$$

where $N$ is the number of internal visual words, $x_{(j)}^I \in \mathbb{R}^D$, $j = 1, \ldots, N$, is the $j$th column of $X^I$, and the subscript of $X_0^I$ denotes the encoding stage. Given $X_0^I$, we model the attraction dependency with the transformer encoding process, which takes the form

$$\begin{align*}
M_1^I &= \phi_i(x_{i-1}^I, x_{i-1}^I, x_{i-1}^I) + x_{i-1}^I, \\
M_1^I &= \gamma(\zeta(M_1^I)) + M_1^I, \quad (2)
\end{align*}$$

where $\gamma(\cdot)$ represents two feed-forward layers, $\zeta(\cdot)$ indicates a linear projection, $i = 1, \ldots, L$, denotes the $i$th encoding stage, and $\phi_i(\cdot)$ denotes the multi-head attention module of the $i$th layer, as shown in Fig. 3. In $\phi_i(\cdot)$, the input visual words are independently projected to a value matrix $V_h$, a key matrix $K_h$, and a query matrix $Q_h$, where $h = 1, \ldots, H$, and are then processed by $H$ attention heads in parallel. For the $h$th head, the attention-weighted matrix $V_h^w \in \mathbb{R}^{D \times N}$ can be computed by

$$V_h^w = \alpha(Q_h K_h^T / \sqrt{d})V_h, \quad (3)$$

where $\alpha(\cdot)$ is the $\text{softmax}$ operator, and $\sqrt{d}$ is a scaling factor [42]. All $V_h^w$’s are concatenated to $\bar{V} \in \mathbb{R}^{HD \times N}$, and $\bar{V}$ is finally projected to the output of the multi-head attention $O \in \mathbb{R}^{D \times N}$.

After $L$ encoding stages in Eq. 2, we acquire the attraction-aware internal visual words

$$X^I = X^I_L = \{x_{(0)}^I, x_{(1)}^I, \ldots, x_{(N)}^I\}, \quad (4)$$

where $x_{(j)}^I \in \mathbb{R}^D$, $j = 1, \ldots, N$, is the $j$th attraction-aware internal visual word.

For the external visual words $X^E \in \mathbb{R}^{D \times N}$, we also model the attraction dependencies between them. Hence, another attraction learner is applied to compute the attraction-aware external visual words $X^E = \{x_{(0)}^E, x_{(1)}^E, \ldots, x_{(N)}^E\}$.

### 3.4. Repulsion Modeling

Modeling attraction can only explain why some visual words concurrently appear, but cannot infer why some words are discarded. Hence, repulsion modeling also matters. To this end, we model repulsion dependencies to understand why some visual words are discarded. As aforementioned, we consider the repulsion dependency to be the semantically/spatially incompatible relation to the main subject. In the context where the main subject is represented as internal visual words, the repulsion dependencies between internal and external visual words could facilitate the evaluation of the composition quality.

The transformer decoder that receives the input of two sequences exactly suits repulsion modeling. Given the attraction-aware internal visual words $X^I$ and the stage-0 external visual words $X^E_0$

$$\begin{align*}
\{X^I = \{x_{(0)}^I, x_{(1)}^I, \ldots, x_{(N)}^I\} \\
X^E_0 = \{x_{(0)}^E, x_{(1)}^E, \ldots, x_{(N)}^E\}\}
\end{align*} \quad (5)$$

the transformer decoder can be formulated by

$$\begin{align*}
M_k^E &= \phi_k'(x_{k-1}^E, x_{k-1}^E, x_{k-1}^E) + x_{k-1}^E, \\
M_k^E &= \gamma(\zeta(M_k^E)) + M_k^E, \\
X_k^E &= \gamma(\zeta(M_k^E)) + M_k^E \quad (6)
\end{align*}$$

where $k = 1, \ldots, M$, denotes the $k$th decoding stage. $\phi_k'(\cdot)$ and $\phi_k''(\cdot)$ are also multi-head attention modules.

After $M$ decoding stages, the repulsion-aware visual words amount to

$$\hat{X}_L^R = \{x_{(0)}^R, x_{(1)}^R, \ldots, x_{(N)}^R\}, \quad (7)$$

where $x_{(j)}^R \in \mathbb{R}^D$, $j = 1, \ldots, N$, is the $j$th repulsion-aware visual word.
3.5. Composition Quality Scoring

The two main goals of TransView are: i) gathering the visual words that are with attractive relation; and ii) discarding the visual words that are repulsive from attractive ones. Therefore, the desired cropping box should trade off between attraction and repulsion. In the fusion of features, we only consider $X^I$ and $X^R$, without $X^E$, to decrease the contribution of external words. Indeed, we observe that $X^E$ has little influence to the final performance. To recover the spatial resolution, $X^I$ and $X^R$ are reshaped into $\hat{X}^I \in \mathbb{R}^{D \times H \times W}$ and $\hat{X}^R \in \mathbb{R}^{D \times H \times W}$, respectively. Finally, $\hat{X}^I$ and $\hat{X}^R$ are concatenated with the original RoI region feature $\mathcal{F}_{RoI}$ as the final attraction-repulsion-aware composition representation. The final feature passes a fully connected layer to predict the score $s$, which will be used to assess the composition quality of a candidate view.

During training, we attempt to force our network to focus on the views with good composition quality, rather than treating all candidates equally. Hence, given the ground truth score $g$, the predicted score $s$ is supervised by a weighted smooth $\ell_1$ loss [27], defined by

$$L = \frac{1}{T} \sum_{t=1}^{T} e^{\frac{\max(0, g_t - \overline{g})}{\delta}} L_1^s(s_t - g_t),$$

where $T$ is the number of the candidate views, $\overline{g}$ is the average score of views in a batch, $\delta$ is a regularization parameter, and $L_1^s$ is the smooth $\ell_1$ loss defined by

$$L_1^s(x) = \begin{cases} 0.5x^2 & \text{if } x < 1 \\ |x| - 0.5 & \text{if } x \geq 1 \end{cases}.$$  

4. Results and Discussion

Here we report and discuss our experimental results. We begin with the used datasets and evaluation metrics.

4.1. Datasets and Evaluation Metrics

Experiments are conducted on the GAIC dataset [52] and FCDB dataset [6]. The GAIC dataset contains 3,336 images split into 2,636 training samples, 200 validation samples, and 500 testing samples. The metrics proposed in [52] are employed for evaluation. They are averaged Pearson correlation coefficient ($\overline{PCC}$), averaged Spearman’s rank-order correlation coefficient ($\overline{SRCC}$), and “return $k$ of top-$n$ accuracy” $\overline{ACC}_{k/n}$. $\overline{PCC}$ and $\overline{SRCC}$ evaluate the ranking consistency between predictions and ground truths. $\overline{ACC}_{k/n}$ measures whether an algorithm can recall the best view. Details of metrics can be found in [52]. The FCDB dataset contains 1743 images with single ground truth cropping box, in which, 1395 images are used for training and 348 images for test. Although the Intersection-over-Union (IoU) metric used in FCDB is not reliable [51], we still report the experimental results on the test set of FCDB to compare with other approaches.

4.2. Implementation Details

The original features generated from MobileNetV2 (pre-trained on ImageNet [10]) are reduced to 32 channels. The aligned size of RoIAlign and RoDAlign is set to $12 \times 12$, which means the length of internal and external visual words is 144. The number of encoding and decoding stages of the transformer equals to $L = M = 6$, the number of heads is set to $H = 4$ in multi-head attention modules, and the scaling factor is $d = 8$. In the training stage, 64 randomly chosen views from one image are batched as the input and the regularization parameter in the weighted smooth $\ell_1$ loss is set to $\delta = 2$. The training samples are resized to ensure that the short side is 256 pixels, and data augmentation strategies follow the same in [52]. The network is optimized by Adam [23] with the learning rate of $5 \times 10^{-5}$ for 100 epochs.

4.3. Performance Comparison

Performance of TransView is compared against other state-of-the-art image cropping and view recommendation methods. For qualitative comparison, we highlight the top-1 cropping view recommendation [7, 51, 52]. Note that, some image cropping approaches [24] that only generate one cropping box per image.

Quantitative Comparison. Quantitative results on the GAIC datasets are reported in Table 1. Note that, A2RL [24] and VPN [48] can only report the metrics of $A2RL_{1/5}$ and $A2RL_{1/10}$ because VPN generates candidate views based on predefined anchors, rather than on-the-fly scoring; and A2RL only generates one cropping box for each image. The performance of CGS [27] are the reported results on a part of the GAIC dataset [51] due to the lack of code. Following the practice of [51], the output boxes of A2RL and VPN are approximated to the nearest candidate views of our method. We observe that, on the GAIC dataset, our TransView significantly outperforms other competitors among the metrics of $A2RL_{1/5}$, $A2RL_{2/5}$ and $A2RL_{1/10}$, which shows that our model can recall the best views more precisely with narrower error bound. On the FCDB dataset, our model also exhibits superior performance compared with the models trained on GAIC dataset. When our model and the models of [51, 52] are not trained on the FCDB, our model shows better generalization.

Qualitative Comparison. Qualitative comparison is shown in Fig. 4. We observe that some methods have notable limitations: i) A2RL cannot crop redundant regions effectively; ii) when VEN and VPN fail to remove redundant...
Figure 4. Qualitative comparison of returned top-1 views. Compared with other methods, our method not only removes the redundancy and preserve the main content more accurately, but also organizes the visual elements in the cropping box more aesthetically.

Table 1. Quantitative comparison to other state-of-the-art approaches on the GAIC dataset [52]. The best performance is in boldface. GAIC∗ indicates the conference version of GAIC.

<table>
<thead>
<tr>
<th>Model</th>
<th>Acc1/5</th>
<th>Acc2/5</th>
<th>Acc3/5</th>
<th>Acc4/5</th>
<th>Acc5</th>
<th>Acc1/10</th>
<th>Acc2/10</th>
<th>Acc3/10</th>
<th>Acc4/10</th>
<th>Acc10</th>
<th>SRCC</th>
<th>PCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>A2RL [24]</td>
<td>23.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>39.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>VPN [48]</td>
<td>36.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>48.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>VFN [7]</td>
<td>26.6</td>
<td>26.5</td>
<td>26.7</td>
<td>25.7</td>
<td>26.4</td>
<td>40.6</td>
<td>40.2</td>
<td>40.3</td>
<td>39.3</td>
<td>40.1</td>
<td>0.485</td>
<td>0.503</td>
</tr>
<tr>
<td>VEN [48]</td>
<td>37.5</td>
<td>35.0</td>
<td>35.3</td>
<td>34.2</td>
<td>35.5</td>
<td>50.5</td>
<td>49.2</td>
<td>48.4</td>
<td>46.4</td>
<td>48.6</td>
<td>0.616</td>
<td>0.662</td>
</tr>
<tr>
<td>GAIC∗ [51]</td>
<td>65.8</td>
<td>61.4</td>
<td>57.6</td>
<td>54.4</td>
<td>62.5</td>
<td>82.4</td>
<td>80.0</td>
<td>78.1</td>
<td>75.6</td>
<td>79.0</td>
<td>0.832</td>
<td>0.857</td>
</tr>
<tr>
<td>GAIC [52]</td>
<td>68.2</td>
<td>65.5</td>
<td>63.0</td>
<td>58.4</td>
<td>63.9</td>
<td>83.0</td>
<td>81.5</td>
<td>78.2</td>
<td>76.0</td>
<td>79.7</td>
<td>0.849</td>
<td>0.874</td>
</tr>
<tr>
<td>CGS [27]</td>
<td>63.0</td>
<td>62.3</td>
<td>58.8</td>
<td>54.9</td>
<td>59.7</td>
<td>81.5</td>
<td>79.5</td>
<td>77.0</td>
<td>73.3</td>
<td>77.8</td>
<td>0.795</td>
<td>-</td>
</tr>
<tr>
<td>TransView</td>
<td>69.0</td>
<td>66.9</td>
<td>61.9</td>
<td>57.8</td>
<td>63.9</td>
<td>85.4</td>
<td>84.1</td>
<td>81.3</td>
<td>78.6</td>
<td>82.4</td>
<td>0.857</td>
<td>0.880</td>
</tr>
</tbody>
</table>

Table 2. Quantitative comparison with other approaches on the FCDB dataset [6]. The models are grouped if they are trained on the GAIC dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>IoU↑</th>
<th>Disp↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>w.o. GAIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A2RL [24]</td>
<td>0.663</td>
<td>0.089</td>
</tr>
<tr>
<td>A3RL [25]</td>
<td>0.696</td>
<td>0.077</td>
</tr>
<tr>
<td>VPN [48]</td>
<td>0.711</td>
<td>0.073</td>
</tr>
<tr>
<td>VEN [48]</td>
<td>0.735</td>
<td>0.072</td>
</tr>
<tr>
<td>ASM [40]</td>
<td>0.749</td>
<td>0.068</td>
</tr>
<tr>
<td>w. GAIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAIC∗ [51]</td>
<td>0.672</td>
<td>0.084</td>
</tr>
<tr>
<td>GAIC [52]</td>
<td>0.673</td>
<td>-</td>
</tr>
<tr>
<td>TransView</td>
<td>0.682</td>
<td>0.080</td>
</tr>
</tbody>
</table>

4.4. Visualization and Analysis

To evaluate the contribution of different components of the proposed model, we conduct ablation studies on the small GAIC dataset [51] to reduce the training cycles. Furthermore, a series of visualizations are illustrated to reveal why our model can outperform the other methods.

Impact of Attraction and Repulsion. The main argument of this work is that dependencies between visual elements matter. Compared with our baseline that predicts the mean opinion scores of the candidate views based on content-aware region features, we further model attraction and repulsion in this work. To explore the impact of attraction and repulsion, different combinations of attraction, repulsion, and content-aware region features are evaluated. Results are listed in Table 3. We can make the following observations:

unnecessary visual elements, which breaks the visual balance. However, our model can precisely place the main elements following the rule of thirds and organize the visual elements in harmony, which demonstrates the efficacy of explicit modeling of the attraction and repulsion.
Figure 5. Qualitative comparison with the baseline GAIC. The green dotted lines indicate the best location for the main elements according to the rule of thirds. The red rectangles box outside the regions violates the visual balance. It is clear that our method produces cropping views that are close to top-1 annotations and that obey real aesthetic composition rules.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>✓</td>
<td></td>
<td></td>
<td>60.5</td>
<td>77.5</td>
<td>0.766</td>
</tr>
<tr>
<td>2</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>64.0</td>
<td>80.5</td>
<td>0.790</td>
</tr>
<tr>
<td>3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>68.5</td>
<td>83.0</td>
<td><strong>0.803</strong></td>
</tr>
<tr>
<td>4</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>66.0</td>
<td>83.0</td>
<td>0.791</td>
</tr>
<tr>
<td>5</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>64.5</td>
<td>82.9</td>
<td>0.788</td>
</tr>
<tr>
<td>6</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>62.0</td>
<td>78.5</td>
<td>0.786</td>
</tr>
</tbody>
</table>

Table 3. Ablation study on attraction and repulsion. Att.: attraction modeling; Rep.: repulsion modeling; Cont.: region-aware RoI features. Best performance is in boldface, and the second best is underlined.

- **Attraction and repulsion are both helpful.** Introducing attraction dependencies or repulsion dependencies can both improve the performance. (No. 4 vs. No. 6 and No. 5 vs. No. 6)

- **Attraction and repulsion are complementary.** When attraction and repulsion are fused, the performance is further enhanced, which suggests that attraction and repulsion benefit each other. (No. 3 vs. No. 4 and No. 3 vs. No. 5)

- **Content information is indispensable.** Performance drops when content-aware region features are discarded. It indicates that content information serves as the foundation of attraction and repulsion. (No. 1 vs. No. 4 and No. 2 vs. No. 3)

**Impact of Positional Embedding.** Akin to the attraction and repulsion between molecules are related to their relative positions, here we justify the role of positional embedding. The baseline is a model without positional embedding. We also compare a trigonometric function-based positional encoding approach. Results in Table 4 illustrate that i) modeling the position of visual elements is of significant importance for attraction and repulsion encoding, and ii) such positional information cannot be simply described by a preset function, which implies the attraction and repulsion between visual elements do relate to positions, but such positions are not immediately interpretable by humans and should be learned in a parametric manner.

**Feature Distribution With Attraction and Repulsion Modeling.** Further experiments are conducted to study why modeling the attraction and repulsion can boost the performance of cropping. One challenging problem of cropping is how to distinguish candidate views share almost the same image content. To discriminate different candidates, the ideal distribution is that each feature used to predict the aesthetic score preserves sufficient margins to others. As shown in Fig. 6, the final features are visualized by t-SNE [41]. For the baseline, the features of similar views tend to cluster closely. This is not desirable because it would be hard to discriminate the composition differences. In other words, clear clusters would imply ambiguous aes-

Figure 6. Comparison of view-wise feature distribution. The feature distribution for candidate views is visualized using t-SNE [41]. The darker the colors are, the higher the rank of a view is. The view-wise features of the baseline show clear clusters, while the features of our model are more separable, which suggests views with visually similar content can be discriminated more clearly in our model.

<table>
<thead>
<tr>
<th>Positional Embedding</th>
<th>Acc1/5</th>
<th>Acc1/10</th>
<th>SRCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learnable</td>
<td>68.5</td>
<td>83.0</td>
<td><strong>0.803</strong></td>
</tr>
<tr>
<td>Trigonometric [42]</td>
<td>46.0</td>
<td>61.5</td>
<td>0.722</td>
</tr>
<tr>
<td>None</td>
<td>45.5</td>
<td>58.0</td>
<td>0.721</td>
</tr>
</tbody>
</table>

Table 4. Ablation study on positional embedding.
Figure 7. **Attention maps of the reference words.** The red and blue points indicate the internal and external visual words. All attention maps visualize the content inside the cropping box only. (a) shows the cropping view of the top-1 prediction, and (b) is the view that ranks the 40th with the truncation phenomenon.

Figure 8. **Attraction and repulsion weights between visual words.** Internal and external visual words are marked by pink and blue boxes, respectively. Red and yellow denote the attraction and repulsion weights, respectively.

Thethetic scores. For the proposed model, the feature distribution is separable, and the margins between features are clear. This reveals that the attraction and repulsion dependencies amplify the differences between candidate views.

**Visualization of Attractive and Repulsive Attention Maps.** We step further to explore how the attraction and repulsion are encoded in the proposed model. The attention maps of the last attention layer of the internal attraction learner and repulsion learner are visualized. The visualized attention maps of the predicted top-1 and top-40 cropping views are compared in Fig. 7, from which we can observe that: i) the internal visual words can focus on the main visual elements of the image without supervision even the cropping box is of poor composition quality. This also explains why the positional embedding is of importance. When internal visual words focus on the main elements, the location of focused elements relative to the cropping box is a strong indication to the assessment of composition quality; ii) the external visual words that should be discarded do not respond actively to the content inside the cropping box, and the words that are incorrectly classified to external words still respond actively to the in-box content. In this way, our model can detect whether a mistake, e.g., the truncation problem, occurs.

To show the pair-wise relation between visual words, the attention weights of internal attraction learner and repulsion learner are grouped and normalized to the range of 0 to 1 for ease of exposition. Two examples are illustrated in Fig. 8. It can be observed that, the attraction weights between internal words are generally greater than 0.5, while the repulsion weights between internal and external words are more likely close to 0, which suggests the transformer indeed models the attraction and repulsion appropriately, even without explicit supervision.

**5. Conclusion**

In this work, we rethink the validity of existing automatic image cropping methods that only represent visual elements by extracting region-based convolutional features. We argue that the presence of visual elements is not sufficient for image cropping or view recommendation, but the relation between elements matters and makes for effective cropping view recommendation. By decomposing the relation into attraction and repulsion, we model these two types of relation by the transformer encoder and decoder, respectively. Extensive experimental results demonstrate the effectiveness of incorporating relation information, and additional analyses reveal the characteristics and soundness of attraction and repulsion between visual elements.

**Acknowledgements.** This work was funded by the DigiX Joint Innovation Center of Huawei-HUST.
References


