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

Zhiyu Pan[1] Zhiguo Cao[1] Kewei Wang[1] Hao Lu[1,*] Weicai Zhong[2]
[1]Key Laboratory of Image Processing and Intelligent Control, Ministry of Education
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
[2]Huawei 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].

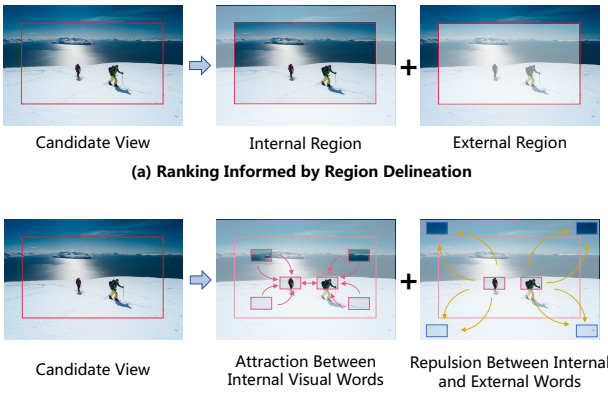

Figure 1. **Conceptual difference between prior arts and ours.** (a) Predecessors recompose images with region-of-interest (RoI) and region-of-discard (RoD) features [51] which depict the presence of visual components rather than the organization. (b) Our insight is to model the organization of visual elements (image patches) using attraction and repulsion dependencies.

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-

*Corresponding author

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 image 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.

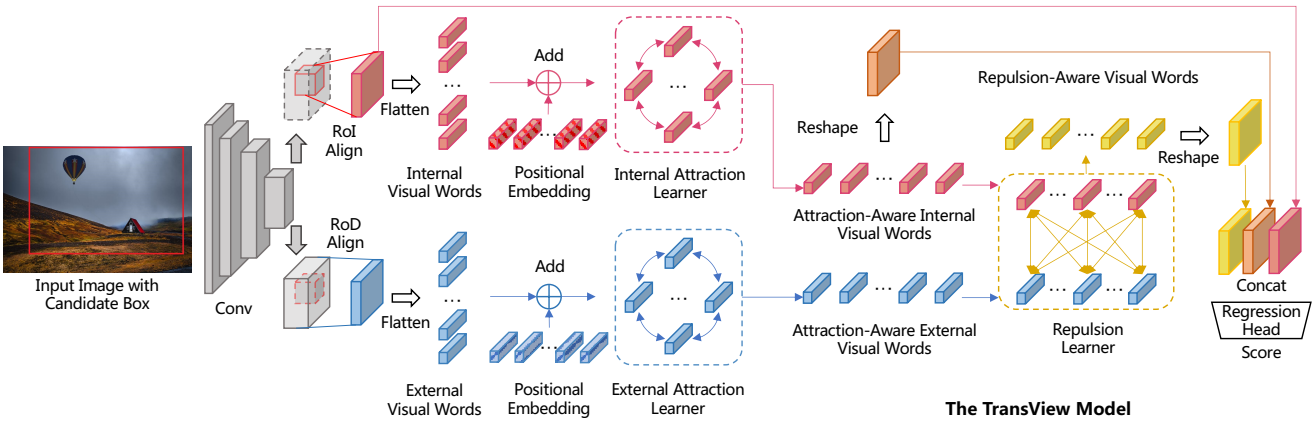

Figure 2. **Technical pipeline of the TransView model.** TransView is composed of a convolutional backbone, two attraction learners, and a repulsion learner. Given a candidate cropping box, the backbone with RoIAlign [16] and RoDAlign [51] generates internal visual words and external visual words to represent the visual elements inside and outside the candidate box. The attraction learners are used to encode the dependencies between visual words of the same regions, and the repulsion learner aims to encode the relation between across-box visual words. The final score is predicted conditioned on three concatenated representations.

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 $\mathcal{F}$ before RoIAlign and RoDAlign. After RoIAlign and RoDAlign, the RoI and RoD features can be extracted and aligned to $\mathcal{F}_{RoI} \in \mathbb{R}^{D\times H\times W}$ and $\mathcal{F}_{RoD} \in \mathbb{R}^{D\times H\times W}$, respectively. By flattening $\mathcal{F}_{RoI}$ and $\mathcal{F}_{RoD}$ into sequences, we can obtain fixed-length internal words $\boldsymbol{C}^I \in \mathbb{R}^{D\times N}$ and external words $\boldsymbol{C}^E \in \mathbb{R}^{D\times N}$, 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 $\boldsymbol{P}^I \in \mathbb{R}^{D\times N}$ and external positional embedding $\boldsymbol{P}^E \in \mathbb{R}^{D\times N}$ 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 $\boldsymbol{X}^I \in \mathbb{R}^{D\times N}$ and external visual words $\boldsymbol{X}^E \in \mathbb{R}^{D\times N}$ by $\boldsymbol{X}^I = \boldsymbol{C}^I + \boldsymbol{P}^I$ and $\boldsymbol{X}^E = \boldsymbol{C}^E + \boldsymbol{P}^E$, 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

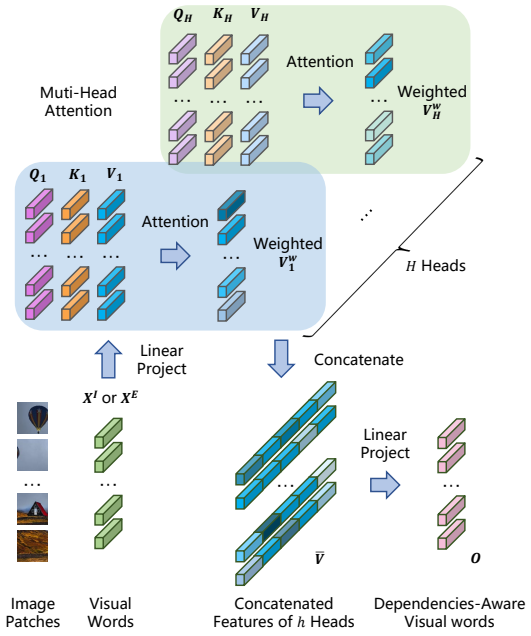

Figure 3. **Multi-head attention.** The visual words $\boldsymbol{X}^I$ or $\boldsymbol{X}^E$ are linearly projected to $\boldsymbol{V}_h$, $\boldsymbol{K}_h$, and $\boldsymbol{Q}_h$ by H times with different projection matrices, where $h = 1, ..., H$. With H parallel attentional operations, the results of $\boldsymbol{V}_0^w, \boldsymbol{V}_1^w, ..., \boldsymbol{V}_H^w$ are concatenated into a matrix $\overline{\boldsymbol{V}} \in \mathbb{R}^{HD\times N}$ and are finally projected into the output $\boldsymbol{O} \in \mathbb{R}^{D\times N}$.

encoder is employed as the attraction learner. The input is a sequence $\boldsymbol{X}_0^I$ formed by internal visual words, defined by

$$\boldsymbol{X}_0^I = \{\boldsymbol{x}_{(0)}^I, \boldsymbol{x}_{(1)}^I, ..., \boldsymbol{x}_{(N)}^I\}, \tag{1}$$

where N is the number of internal visual words, $\boldsymbol{x}_{(j)}^I \in \mathbb{R}^D, j = 1, ..., N$, is the jth column of $\boldsymbol{X}^I$, and the subscript of $\boldsymbol{X}_0^I$ denotes the encoding stage. Given $\boldsymbol{X}_0^I$, we model the attraction dependency with the transformer encoding process, which takes the form

$$\begin{cases} \boldsymbol{M}_i^I = \phi_i(\boldsymbol{X}_{i-1}^I, \boldsymbol{X}_{i-1}^I, \boldsymbol{X}_{i-1}^I) + \boldsymbol{X}_{i-1}^I \\ \boldsymbol{X}_i^I = \gamma(\zeta(\boldsymbol{M}_i^I)) + \boldsymbol{M}_i^I \end{cases}, \tag{2}$$

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

$$\boldsymbol{V}_h^w = \alpha(\frac{\boldsymbol{Q}_h \boldsymbol{K}_h^T}{\sqrt{d}})\boldsymbol{V}_h, \tag{3}$$

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

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

$$\dot{\boldsymbol{X}}^I = \boldsymbol{X}_L^I = \{\dot{\boldsymbol{x}}_{(0)}^I, \dot{\boldsymbol{x}}_{(1)}^I, ..., \dot{\boldsymbol{x}}_{(N)}^I\}, \tag{4}$$

where $\dot{\boldsymbol{x}}_{(j)}^I \in \mathbb{R}^D, j = 1, ..., N$, is the jth attraction-aware internal visual word.

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

3.4. Repulsion Modeling

Modeling attraction can only explain why some visual words concurrently appear, but cannot inform 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 $\dot{\boldsymbol{X}}^I$ and the stage-0 external visual words $\dot{\boldsymbol{X}}_0^E$

$$\begin{cases} \dot{\boldsymbol{X}}^I = \{\dot{\boldsymbol{x}}_{(0)}^I, \dot{\boldsymbol{x}}_{(1)}^I, ..., \dot{\boldsymbol{x}}_{(N)}^I\} \\ \dot{\boldsymbol{X}}_0^E = \{\dot{\boldsymbol{x}}_{(0)}^E, \dot{\boldsymbol{x}}_{(1)}^E, ..., \dot{\boldsymbol{x}}_{(N)}^E\} \end{cases}, \tag{5}$$

the transformer decoder can be formulated by

$$\begin{cases} \boldsymbol{M}_k^E = \phi_k'(\dot{\boldsymbol{X}}_{k-1}^E, \dot{\boldsymbol{X}}_{k-1}^E, \dot{\boldsymbol{X}}_{k-1}^E) + \dot{\boldsymbol{X}}_{k-1}^E \\ \boldsymbol{M}_k^R = \phi_k''(\dot{\boldsymbol{X}}^I, \dot{\boldsymbol{X}}^I, \boldsymbol{M}_k^E) + \boldsymbol{M}_k^E \\ \dot{\boldsymbol{X}}_k^E = \gamma(\zeta(\boldsymbol{M}_k^R)) + \boldsymbol{M}_k^R \end{cases}, \tag{6}$$

where $k = 1, ..., M$, denotes the kth 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

$$\ddot{\boldsymbol{X}}^R = \dot{\boldsymbol{X}}_L^E = \{\ddot{\boldsymbol{x}}_{(0)}^R, \ddot{\boldsymbol{x}}_{(1)}^R, ..., \ddot{\boldsymbol{x}}_{(N)}^R\}, \tag{7}$$

where $\ddot{\boldsymbol{x}}_{(j)}^R \in \mathbb{R}^D, j = 1, ..., N$, is the jth 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 $\dot{\boldsymbol{X}}^I$ and $\ddot{\boldsymbol{X}}^R$, without $\dot{\boldsymbol{X}}^E$, to decrease the contribution of external words. Indeed, we observe that $\dot{\boldsymbol{X}}^E$ has little influence to the final performance. To recover the spatial resolution, $\dot{\boldsymbol{X}}^I$ and $\ddot{\boldsymbol{X}}^R$ are reshaped into $\dot{\mathcal{X}}^I \in \mathbb{R}^{D\times H\times W}$ and $\ddot{\mathcal{X}}^{\mathcal{R}} \in \mathbb{R}^{D\times H\times W}$, respectively. Finally, $\dot{\mathcal{X}}^I$ and $\ddot{\mathcal{X}}^{\mathcal{R}}$ 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 ℓ_1 loss [27], defined by

$$L = \frac{1}{T}\sum_{t=1}^{T} e^{\frac{\max(0, g_t - \overline{g})}{\delta}} L_1^s(s_t - g_t)\,, \tag{8}$$

where T is the number of the candidate views, $\overline{g}$ is the average score of views in a batch, δ is a regularization parameter, and L_1^s is the smooth ℓ_1 loss defined by

$$L_1^s(x) = \begin{cases} 0.5x^2 & \text{if } x < 1 \\ |x| - 0.5 & \text{if } x >= 1 \end{cases}. \tag{9}$$

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" $ACC_{k/n}$. $\overline{PCC}$ and $\overline{SRCC}$ evaluate the ranking consistency between predictions and ground truths. $ACC_{k/n}$ 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 12×12, 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 ℓ_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 5×10^{-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 $Acc_{1/5}$ and $Acc_{1/10}$ 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 $Acc_{1/5}$, $Acc_{2/5}$ and $Acc_{*/10}$, 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 redun-

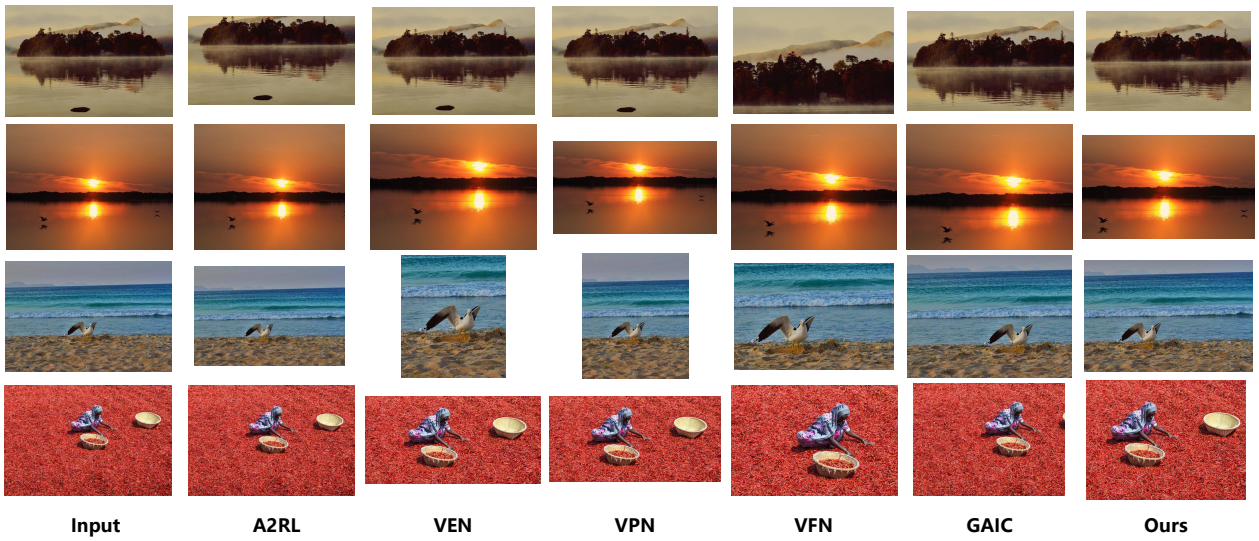

Figure 4. **Qualitative comparison of returned top-1 views.** Compared with other methods, our method not only removes the redundancy and preserve the main content more accurately, but also organizes the visual elements in the cropping box more aesthetically.

Model	$Acc_{1/5}$	$Acc_{2/5}$	$Acc_{3/5}$	$Acc_{4/5}$	$\overline{Acc_5}$	$Acc_{1/10}$	$Acc_{2/10}$	$Acc_{3/10}$	$Acc_{4/10}$	$\overline{Acc_{10}}$	$\overline{SRCC}$	$\overline{PCC}$
A2RL [24]	23.2	-	-	-	-	39.5	-	-	-	-	-	-
VPN [48]	36.0	-	-	-	-	48.5	-	-	-	-	-	-
VFN [7]	26.6	26.5	26.7	25.7	26.4	40.6	40.2	40.3	39.3	40.1	0.485	0.503
VEN [48]	37.5	35.0	35.3	34.2	35.5	50.5	49.2	48.4	46.4	48.6	0.616	0.662
GAIC* [51]	65.8	61.4	57.6	54.4	62.5	82.4	80.0	78.1	75.6	79.0	0.832	0.857
GAIC [52]	68.2	65.5	**63.0**	**58.4**	**63.9**	83.0	81.5	78.2	76.0	79.7	0.849	0.874
CGS [27]	63.0	62.3	58.8	54.9	59.7	81.5	79.5	77.0	73.3	77.8	0.795	-
TransView	**69.0**	**66.9**	61.9	57.8	**63.9**	**85.4**	**84.1**	**81.3**	**78.6**	**82.4**	**0.857**	**0.880**

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.

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

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

dancy, they prefer to maximize the dominance of visual elements without considering aesthetics; iii) VFN tends to focus on irrelevant image content; iv) as the baseline of our work, GAIC [52] generates acceptable results but the visual elements are not organized following an aesthetic rule; v) in contrast, TransView can organize the visual elements exactly in line with principles of art after redundancy removal and the preservation of main elements. Our argument is made obvious in the qualitative comparison between GAIC and ours in Fig. 5 where the main visual elements in the results of baseline shift from the desired position according to the rule of thirds. In addition, the baseline includes 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.

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:

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.

No.	Att.	Rep.	Cont.	$Acc_{1/5}$	$Acc_{1/10}$	$\overline{SRCC}$
1	✓			60.5	77.5	0.766
2	✓	✓		64.0	80.5	0.790
3	✓	✓	✓	**68.5**	**83.0**	**0.803**
4	✓		✓	66.0	**83.0**	0.791
5		✓	✓	64.5	82.9	0.788
6			✓	62.0	78.5	0.786

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

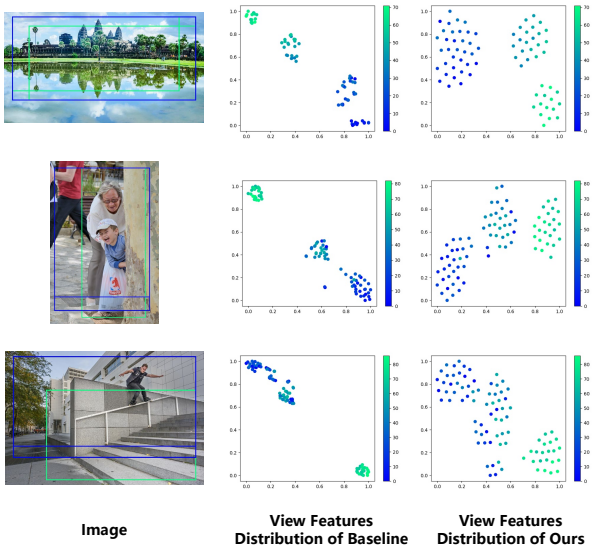

Figure 6. **Comparison of view-wise feature distribution.** The feature distribution for candidate views is visualized using t-SNE [41]. The darker the colors are, the higher the rank of a view is. The view-wise features of the baseline show clear clusters, while the features of our model are more separable,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

Positional Embedding	$Acc_{1/5}$	$Acc_{1/10}$	$\overline{SRCC}$
Learnable	68.5	83.0	0.803
Trigonometric [42]	46.0	61.5	0.722
None	45.5	58.0	0.721

Table 4. Ablation study on positional embedding.

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 descried 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-

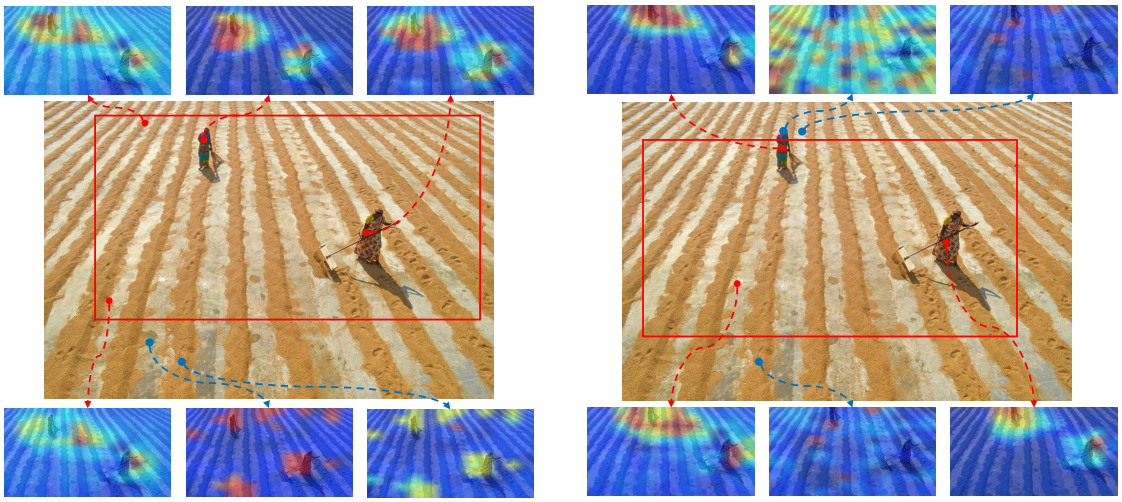

Figure 7. **Attention maps of the reference words.** The red and blue points indicate the internal and external visual words. All attention maps visualize the content inside the cropping box only. (a) shows the cropping view of the top-1 prediction, and (b) is the view that ranks the 40th with the truncation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

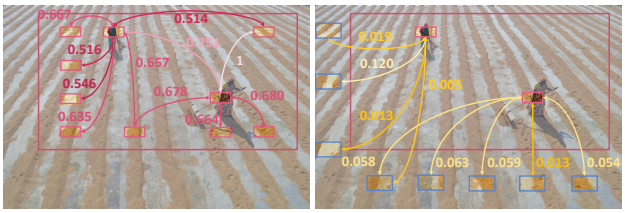

Figure 8. **Attraction and repulsion weights between visual words.** Internal and external visual words are marked by pink and blue boxes, respectively. Red and yellow denote the attraction and repulsion weights, respectively.

thetic 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

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proc. Int. Conf. Learn. Represent.*, 2015. 2

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Eur. Conf. Comput. Vis.*, pages 213–229, 2020. 2, 3

[3] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12299–12310, 2021. 2

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

[5] Xiangrong Chen and HongJiang Zhang. Text area detection from video frames. In *Pacific-Rim Conf. Multimedia*, pages 222–228, 2001. 2

[6] Yi-Ling Chen, Tzu-Wei Huang, Kai-Han Chang, Yu-Chen Tsai, Hwann-Tzong Chen, and Bing-Yu Chen. Quantitative analysis of automatic image cropping algorithms: A dataset and comparative study. In *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, pages 226–234, 2017. 5, 6

[7] Yi-Ling Chen, Jan Klopp, Min Sun, Shao-Yi Chien, and Kwan-Liu Ma. Learning to compose with professional photographs on the web. In *Proc. ACM Int. Conf. Multimedia*, pages 37–45, 2017. 2, 5, 6

[8] Bin Cheng, Bingbing Ni, Shuicheng Yan, and Qi Tian. Learning to photograph. In *Proc. ACM Int. Conf. Multimedia*, pages 291–300, 2010. 1, 2

[9] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 2

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 248–255, 2009. 5

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. Int. Conf. Learn. Represent.*, 2020. 2

[12] Bernard Dunstan. *Composing your paintings*. Watson-Guptill Publications, 1971. 2, 3

[13] Chen Fang, Zhe Lin, Radomir Mech, and Xiaohui Shen. Automatic image cropping using visual composition, boundary simplicity and content preservation models. In *Proc. ACM Int. Conf. Multimedia*, pages 1105–1108, 2014. 2

[14] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *Proc. Int. Conf. Mach. Learn.*, pages 1243–1252, 2017. 3

[15] Luca Greco and Marco La Cascia. Saliency based aesthetic cut of digital images. In *Proc. Int. Conf. Image Anal. Process.*, pages 151–160, 2013. 2

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proc. Int. Conf. Comput. Vis.*, pages 2961–2969, 2017. 3

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 2

[18] Chaoyi Hong, Shuaiyuan Du, Ke Xian, Hao Lu, Zhiguo Cao, and Weicai Zhong. Composing photos like a photographer. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7057–7066, 2021. 1

[19] Lin Huang, Jianchao Tan, Ji Liu, and Junsong Yuan. Hand-transformer: Non-autoregressive structured modeling for 3d hand pose estimation. In *Eur. Conf. Comput. Vis.*, pages 17–33, 2020. 2

[20] Md Baharul Islam, Wong Lai-Kuan, and Wong Chee-Onn. A survey of aesthetics-driven image recomposition. *Mutltimedia Tools Appl.*, 76(7):9517–9542, 2017. 1

[21] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1072–1080, 2015. 2

[22] Yoon Kim, Carl Denton, Luong Hoang, and Alexander M Rush. Structured attention networks. In *Proc. Int. Conf. Learn. Represent.*, 2017. 2

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[24] Debang Li, Huikai Wu, Junge Zhang, and Kaiqi Huang. A2-rl: Aesthetics aware reinforcement learning for image cropping. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8193–8201, 2018. 2, 5, 6

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

[26] Debang Li, Junge Zhang, and Kaiqi Huang. Learning to learn cropping models for different aspect ratio requirements. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12685–12694, 2020. 2

[27] Debang Li, Junge Zhang, Kaiqi Huang, and Ming-Hsuan Yang. Composing good shots by exploiting mutual relations. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4213–4222, 2020. 1, 2, 5, 6

[28] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. Index networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1–1, 2020. 2

[29] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Proc. Adv. Neural Inform. Process. Syst.*, volume 29, pages 4905–4913, 2016. 2

[30] Luca Marchesotti, Claudio Cifarelli, and Gabriela Csurka. A framework for visual saliency detection with applications to image thumbnailing. In *Proc. Int. Conf. Comput. Vis.*, pages 2232–2239, 2009. 2

[31] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2408–2415, 2012. 1

[32] Masashi Nishiyama, Takahiro Okabe, Yoichi Sato, and Imari Sato. Sensation-based photo cropping. In *Proc. ACM Int. Conf. Multimedia*, pages 669–672, 2009. 2

[33] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *Proc. Int. Conf. Mach. Learn.*, pages 4055–4064, 2018. 2

[34] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4510–4520, 2018. 3

[35] Anthony Santella, Maneesh Agrawala, Doug DeCarlo, David Salesin, and Michael Cohen. Gaze-based interaction for semi-automatic photo cropping. In *In Proc. SIGCHI Conf. Hum. Fact. Comput. Syst.*, pages 771–780, 2006. 2

[36] Jürgen Schmidhuber and Sepp Hochreiter. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997. 2

[37] Henry Schneiderman and Takeo Kanade. A statistical method for 3d object detection applied to faces and cars. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 746–751, 2000. 2

[38] Bongwon Suh, Haibin Ling, Benjamin B Bederson, and David W Jacobs. Automatic thumbnail cropping and its effectiveness. In *Proc. Annu. ACM Symp. User Interface Softw. Technol.*, pages 95–104, 2003. 1, 2

[39] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Proc. Adv. Neural Inform. Process. Syst.*, volume 27, page 3104–3112. Curran Associates, Inc., 2014. 2

[40] Yi Tu, Li Niu, Weijie Zhao, Dawei Cheng, and Liqing Zhang. Image cropping with composition and saliency aware aesthetic score map. In *Proc. of AAAI Conf. Artif. Intell.*, volume 34, pages 12104–12111, 2020. 2, 6

[41] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *J. Mach. Learn. Res.*, 9(11), 2008. 7

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Proc. Adv. Neural Inform. Process. Syst.*, volume 30, pages 6000–6010, 2017. 2, 3, 4, 7

[43] Eleonora Vig, Michael Dorr, and David Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2798–2805, 2014. 2

[44] Wenguan Wang and Jianbing Shen. Deep cropping via attention box prediction and aesthetics assessment. In *Proc. Int. Conf. Comput. Vis.*, pages 2186–2194, 2017. 2

[45] Wenguan Wang, Jianbing Shen, and Haibin Ling. A deep network solution for attention and aesthetics aware photo cropping. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(7):1531–1544, 2018. 2

[46] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7794–7803, 2018. 2

[47] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8741–8750, 2021. 2

[48] Zijun Wei, Jianming Zhang, Xiaohui Shen, Zhe Lin, Radomir Mech, Minh Hoai, and Dimitris Samaras. Good view hunting: Learning photo composition from dense view pairs. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5437–5446, 2018. 1, 2, 5, 6

[49] Wikipedia contributors. Composition (visual arts) — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Composition_(visual_arts), 2021. [Online; accessed 1-March-2021]. 1

[50] Jianzhou Yan, Stephen Lin, Sing Bing Kang, and Xiaoou Tang. Learning the change for automatic image cropping. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 971–978, 2013. 2

[51] Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Reliable and efficient image cropping: A grid anchor based approach. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5949–5957, 2019. 1, 2, 3, 5, 6

[52] Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Grid anchor based image cropping: A new benchmark and an efficient model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020. 2, 3, 5, 6

[53] Mingju Zhang, Lei Zhang, Yanfeng Sun, Lin Feng, and Weiying Ma. Auto cropping for digital photographs. In *Proc. IEEE Int. Conf. Multimedia Expo*, pages 4–pp, 2005. 2

[54] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2