Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Full Length Article

Aesthetic image cropping meets VLP: Enhancing good while reducing bad[☆],[☆☆]

Quan Yuan, Leida Li, Pengfei Chen[*]

School of Artificial Intelligence, Xidian University, Xi'an, Shaanxi, 210044, China

ARTICLE INFO

Keywords:
Image cropping
Vision–Language Pre-training
Aesthetic quality assessment

ABSTRACT

Aesthetic Image Cropping (AIC) enhances the visual appeal of an image by adjusting its composition and aesthetic elements. People make these adjustments based on these elements, aiming to enhance appealing aspects while minimizing detrimental factors. Motivated by these observations, we propose a novel approach called CLIPCropping, which simulates the human decision-making process in AIC. CLIPCropping leverages Contrastive Language–Image Pre-training (CLIP) to align visual perception with textual description. It consists of three branches: composition embedding, aesthetic embedding, and image cropping. The composition embedding branch learns principles based on Composition Knowledge Embedding (CKE), while the aesthetic embedding branch learns principles based on Aesthetic Knowledge Embedding (AKE). The image cropping branch evaluates the quality of candidate crops by aggregating knowledge from CKE and AKE; an MLP produces the best result. Extensive experiments on three benchmark datasets — GAICD-1236, GAICD-3336, and FCDB — show that CLIPCropping outperforms state-of-the-art methods and provides insightful interpretations.

1. Introduction

The emergence of social media platforms has elevated visual content to the forefront of online communication, where compelling images can resonate with viewers and leave a lasting impression. As the demand for high-quality and aesthetically pleasing images increases, mastering elements like composition, lighting, and color theory [1–3] has become essential, although challenging, especially for amateur photographers. Aesthetic Image Cropping (AIC) has attracted attention due to its potential to enhance image composition and visual appeal in various domains, including aesthetic quality assessment [4–11], image enhancement [12,13], image harmonization [14], image generation [15], image retargeting [16,17] and image customization [18,19].

In the literature, several AIC models have been reported. Early efforts focused on designing hand-crafted features based on saliency [20,21], or color [22] that could characterize significant editing before and after cropping. Although hand-crafted features have good physical meanings, they do not align with human perception. This is because human perception during cropping is usually highly abstract and complex.

Recently, Vision–Language Pre-training (VLP) has demonstrated its advantages in a wide range of tasks, such as image segmentation [23,24], image generation [25,26], image captioning [27,28], person ReID [29,30], image denoising [31], and image super-resolution [32]. However, how to take advantage of VLP in the AIC domain and obtain cropping results with interpretability has yet to be explored. Typically, people perform AIC task considering a series of visual elements. As illustrated in Fig. 1, during the cropping process, people first analyze the composition and visual elements of an image, and then enhance the good visual elements during the cropping process. However, many existing AIC models usually simply map the candidate crops to a latent cropping feature space, which makes the cropping results inconsistent with human perception. Furthermore, the absence of interpretability could impede their deployment in practical settings.

Motivated by the above facts, this paper presents a CLIP-based aesthetic image cropping model, dubbed CLIPCropping. When people perform the AIC task, they usually focus on the cropping elements of the image that affect the aesthetics. Composition Knowledge Embedding (CKE) is introduced to extract features of composition elements, and Aesthetic Knowledge Embedding (AKE) is introduced to extract the features of aesthetic elements during the cropping process. Subsequently, these features are combined to obtain cropping features

[☆] This paper has been recommended for acceptance by Zicheng Liu.

[☆☆] The authors thank Yipo Huang (XDU & NTU), Yu Zhou (CUMT) and Xiangfei Sheng (XDU) for the insightful discussions. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants 62471349, 62171340, 61991451, and 62301378.

[*] Corresponding author.
E-mail addresses: dylan.yuanquan@stu.xidian.edu.cn (Q. Yuan), ldli@xidian.edu.cn (L. Li), chenpengfei@xidian.edu.cn (P. Chen).

https://doi.org/10.1016/j.jvcir.2024.104316
Received 25 June 2024; Received in revised form 27 September 2024; Accepted 13 October 2024
Available online 18 October 2024

1047-3203/© 2024 Elsevier Inc. All rights reserved, including those for text and data mining, AI training, and similar technologies.

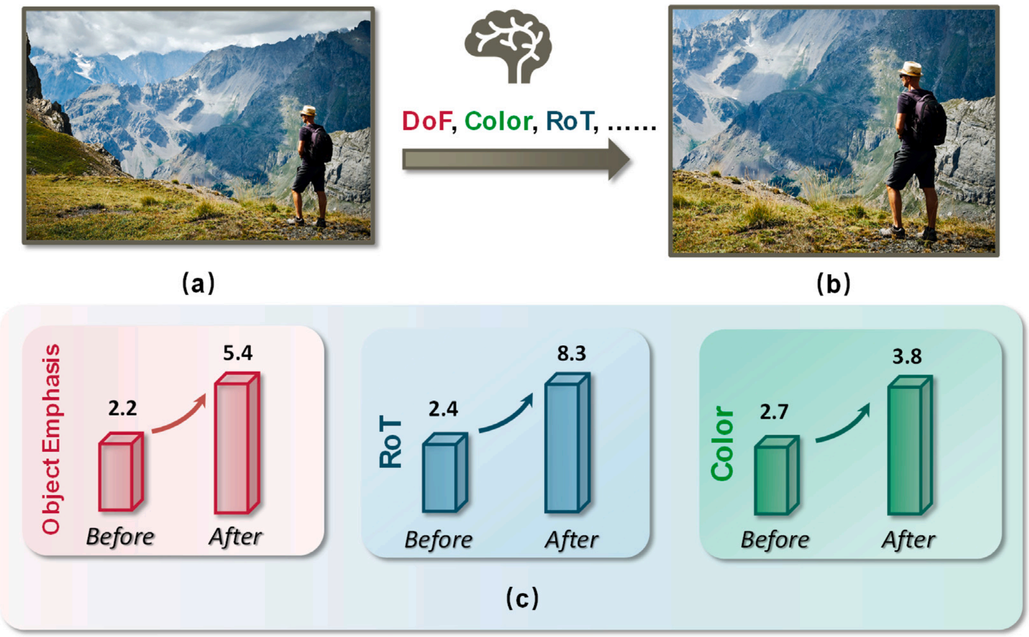


Fig. 1. Motivation behind CLIPCropping. **(a)** original image, **(b)** crop result of CLIPCropping, and **(c)** the change of visual elements before and after cropping. People evaluate and adjust images based on established composition rules and aesthetic elements. Three elements have been enhanced for a more appealing aesthetic result in this example, including object emphasis, rule of thirds, and color, as shown in **(c)**.

incorporating aesthetic and compositional characteristics. Then, these features are fused to generate the enhanced cropping features. Finally, a cropping regressor is introduced to evaluate the different candidate crops, producing in the final cropping result.

In practical applications, CLIPCropping has the potential to offer advantages in industries such as social media, photography, and graphic design. On social media platforms, where high-quality visual content is important for user engagement, CLIPCropping can help automatically enhance the aesthetic composition of images, making them more appealing to viewers. In photography, particularly for amateur photographers, this approach can reduce the need for manual adjustments by aligning the cropped images with aesthetic and compositional principles, thus streamlining workflows. Likewise, in graphic design, CLIPCropping can assist designers in achieving optimal compositions more efficiently by leveraging its ability to align visual perception with textual descriptions. This alignment helps ensure that the cropping decisions not only reflect human aesthetic judgment but are also interpretable, making the tool potentially valuable in real-world settings.

The contributions of this paper can be summarized in three aspects:

- We propose a vision–language-based image cropping method, dubbed CLIPCropping, which leverages rich prior knowledge of composition and aesthetic attributes to facilitate image cropping.
- We propose composition knowledge embedding and aesthetic knowledge embedding, which are designed to model the changes in composition and aesthetic elements that are essential to the AIC task. They work collaboratively to provide rich prior information on image aesthetics and guide the cropping process.
- We conduct extensive experiments and comparisons on three benchmark datasets, including GAICD-3336, GAICD-1236, and FCDB. The results demonstrate the advantages of CLIPCropping, showing that it not only outperforms existing methods but also aligns with human intuition.

2. Related work

This section provides an overview of the research in Image Cropping and Vision–Language Pre-training, which are highly pertinent to our study.

2.1. Image cropping

We review existing image cropping as score evaluation-based methods and regression-based methods. Evaluation-based methods mainly focus on assigning a cropping score to different candidate crops in an image. The score is typically obtained through modeling cropping principles. Chen et al. [33] proposed a model with pairwise ranking constraints, which can distinguish the relative quality of two candidate crops. Wei et al. [34] proposed a cropping method via knowledge transfer, which can obviate the label assignment procedure. Li et al. [35] proposed a graph-based module with a gated feature, which can model the relationship between different candidate crops through graph learning. Tu et al. [36] proposed a composition and saliency-based method, which evaluates candidate crops with an aesthetic map and places salient objects in a proper position. Pan et al. [37] proposed a transformer-based solution that represents visual elements as visual words and models the dependencies between visual words to model the relation between inside and outside the cropping bounding box. Zhang et al. [38] proposed a reinforcement learning-based method that collaborates with an emotion attention-aware map to predict the optimal candidate crops. Zeng et al. [39] proposed a novel interpolation method designed specifically for image cropping, which can model the spatial information. Ni et al. [40] proposed a module embedded with composition knowledge, which can adaptively explore suitable composition rules for the images in a direct and interpretable way. Zhang et al. [41] proposed a human-centered method to handle portrait images with partition-aware and content-preserving features. Wang et al. [42] proposed a model with spatial-aware features and a pairwise classifier, which takes both spatial information and the dominant object into consideration.

Regression-based methods generally work by directly predicting the coordinates of the best crop. Suh et al. [20] used a saliency map and human face as prior knowledge to evaluate candidate crops. Li et al. [43] proposed a reinforcement learning model to provide coordinate crops with a flexible aspect ratio. Lu et al. [44] proposed a weakly supervised method based on aesthetic distributions. Hong et al. [45] proposed a saliency-guided model, which encodes several composition rules in the cropping process. Jia et al. [46]

regarded AIC as a set prediction problem, where multiple crops with a validity classifier were used to match diverse good crops.

In summary, evaluation-based methods are essentially a score regression process, so generally, they can achieve faster inference speed. The limitation is that they can only provide pre-defined cropping results and cannot meet diverse cropping needs. Regression-based methods consider the cropping task as a coordinate regression problem, so that they can meet diverse cropping requirements, such as different aspect ratios. However, typically based on single visual modality and they lack interpretability, which is highly desired in real-world applications.

2.2. Vision–Language Pre-training

Vision–Language Pre-training (VLP) has become a hot research domain, aspiring to craft models that seamlessly integrate the understanding of images and text. These models harness extensive datasets of image–text pairs to forge a conjoint representation, aligning the semantic essence of both modalities [47]. The forerunners in this field, like VilBert [48] and LXMERT [49], pioneered the adaptation of BERT-like architectures to accommodate visual and textual inputs. These initial endeavors paved the way for subsequent innovations like KVL-Bert [50], which incorporated knowledge vectors into the pretraining regimen. The advent of CLIP (Contrastive Language–Image Pre-training) by Radford et al. [51] marked a significant leap in VLP. Drawing on contrastive learning principles similar to those in SimCLR [52], CLIP-aligned images and text within a shared embedding space enabling it to perform tasks zero-shot task without task-specific fine-tuning. Trained on a diverse internet-sourced image–text dataset, CLIP has acquired a broad understanding of various classes and concepts reminiscent of robust representation learning. CLIP's architecture is influenced by multimodal predecessors like ViLBERT by Lu et al. [48]. The unified VLP framework of Zhou et al. [53] and the amalgamation of visual and linguistic representations in VL-BERT by Su et al. [54] parallel CLIP's goals. The alignment of object semantics with pre-training proposed by Li et al. [55] has been fundamental to CLIP's modality harmonization approach. Raffel et al. [56] opens the potential of applying a transformer to CLIP, which becomes a central design of the CLIP family. Similarly, Socratic Models by Zellers et al. [57] embodied the vision of utilizing multimodal data for reasoning, also embraced by CLIP.

CLIP's prominence in the VLP landscape is a testament to the cumulative advancements in the field. It synthesizes various concepts and methodologies from past research, highlighting the efficacy of contrastive learning and the importance of diverse training data for a comprehensive grasp of visual and textual content. Despite these strides, the specific application of CLIP to fields like AIC still needs to be expanded.

3. Proposed model

The overall structure of the proposed CLIPCropping is shown in Fig. 2. Specifically, inspired by the human processing of image cropping, we first construct two knowledge extractors to extract compositional and aesthetic knowledge based on Composition Knowledge Embedding (CKE) and Aesthetic Knowledge Embedding (AKE), respectively. Then, these two features are fused to obtain the enhanced cropping features. Finally, the best cropping result is accessed through an MLP.

3.1. Composition embedding

Hoh et al. [58] demonstrated that compositional attributes, such as spatial positioning, play a critical role in image aesthetics. Therefore, CLIPCropping incorporates a dedicated branch to model compositional features. Specifically, we introduce a CNN module in the composition embedding branch, which learns compositional knowledge using a set of designed prompts $\mathbf{P}_{\alpha}^{comp}$ representing five classical composition rules: center composition, rule of thirds composition, horizontal composition, vertical composition, and curved composition, as shown in Fig. 2-(a). These compositional patterns effectively capture the differences before and after cropping. Center and rule of thirds compositions typically involve changes in the visual subject during cropping, while horizontal composition emphasizes balance along the horizontal axis, vertical composition highlights the visual flow along the vertical axis, and curved composition captures dynamic lines and flow within the image. This compositional knowledge is embedded into the CKE module $\mathbb{F}_{\text{CKE}}$ and $\mathbb{F}_{\text{CLIP-visual}}$ through a contrastive learning approach. Since the CADB [59] dataset, which used for extract composition knowledge, provides boolean labels, images labeled as 0 are considered negative samples for the corresponding composition prompt, while those labeled as 1 are treated as positive samples.

The CKE module consists of multiple linear layers and nonlinear activation functions, and integrates an AttentionPool2d module to facilitate efficient feature aggregation, allowing the model to effectively capture and fuse compositional knowledge.

In optimizing $\mathbb{F}_{\text{CLIP-visual}}$ and $\mathbb{F}_{\text{CKE}}$, we use a subset of the CADB [59] $\mathcal{D}_{comp} = \{\boldsymbol{x}_{comp}^{i}, \boldsymbol{y}_{comp}^{i}\}_{i=1}^{N_{comp}}$, which provides the composition images $\boldsymbol{x}_{comp}$ and their associated five compositional attributes $\boldsymbol{y}_{comp}$ which are mentioned above.

Motivated by CoCoOP [60], the Composition Embedding is optimized through InfoNCE [61]:

$$\mathcal{L}_{\text{comp}}(\phi_{\text{comp}}^{visual}, \phi_{\text{comp}}^{textual}; \Theta_{\text{comp}}) = -\log \frac{\exp(\text{sim}(\phi_{\text{comp}}^{visual}, \phi_{\text{comp}}^{textual+}))}{\sum_{j=0}^{5} \exp(\text{sim}(\phi_{\text{comp}}^{visual}, \phi_{\text{comp},j}^{textual}))}, \tag{1}$$

where $\Theta_{\text{comp}} = \{\theta_{\text{CKE}}, \theta_{\text{CLIP}}, \theta_{\text{learner}}^{comp}\}$, $\phi_{\text{comp}}^{textual}, \phi_{\text{comp}}^{textual+}, \phi_{\text{comp}}^{visual}$ is the textual composition prompt feature, the ground truth composition prompt feature, and composition visual feature, respectively. $\phi_{\text{comp}}^{visual}$ is obtained through $\mathbb{F}_{\text{CLIP-visual}}$ and $\mathbb{F}_{\text{CKE}}$ with a composition image $\boldsymbol{x}_{\text{comp}}$:

$$\phi_{\text{comp}}^{\text{visual}} = \mathbb{F}_{\text{CKE}}(\mathbb{F}_{\text{CLIP-Visual}}(\boldsymbol{x}_{\text{comp}})). \tag{2}$$

and composition textual feature $\phi_{\text{comp}}^{textual}$ is obtained through textual encoder $\mathbb{F}_{\text{textEnc}}$:

$$\phi_{\text{comp}}^{textual} = \mathbb{F}_{\text{texEnc}}(\mathbf{P}_{\beta}^{comp}, \sigma_{embed}^{comp}), \tag{3}$$

where learned composition prompts $\mathbf{P}_{\beta}^{comp}$ is obtained through CLIP tokenizer $\mathbb{F}_{\text{CLIP-tokenizer}}$:

$$\mathbf{P}_{\beta}^{comp} = \mathbb{F}_{\text{CLIP-tokenizer}}(\mathbf{P}_{\alpha}^{comp}), \tag{4}$$

and composition visual embedding σ_{embed}^{comp} is obtained through the composition prompt learner $\mathbb{F}_{learner}^{comp}$ with the composition visual feature $\phi_{\text{comp}}^{\text{visual}}$:

$$\sigma_{embed}^{comp} = \mathbb{F}_{learner}^{comp}(\phi_{\text{comp}}^{\text{visual}}). \tag{5}$$

3.2. Aesthetic embedding

Aesthetic properties are equally crucial for the AIC task. Therefore, in CLIPCropping, a dedicated branch is designed to model aesthetic features, as illustrated in Fig. 2-b. Specifically, we introduce a CNN module, structurally consistent with the CKE, referred to as AKE. This module learns aesthetic knowledge through a set of designed prompts $\mathbf{P}^{aes}$ representing five core aesthetic elements: color, lighting, depth of field, well-emphasized object, and interesting content. These aesthetic elements are essential for capturing the visual appeal of an image. Notably, changes in some aesthetic elements, such as color and lighting, can result from alterations in the image structure due to cropping. In contrast, other elements, such as depth of field, emphasize the focus and sharpness of the image; well-emphasized objects highlight key visual subjects; and interesting content ensures that the image conveys an engaging narrative or concept. This aesthetic knowledge is embedded into the AKE module $\mathbb{F}_{\text{AKE}}$ and $\mathbb{F}_{\text{CLIP-visual}}$ through a contrastive learning approach. Since the AADB [62] dataset, which used for extract aesthetic

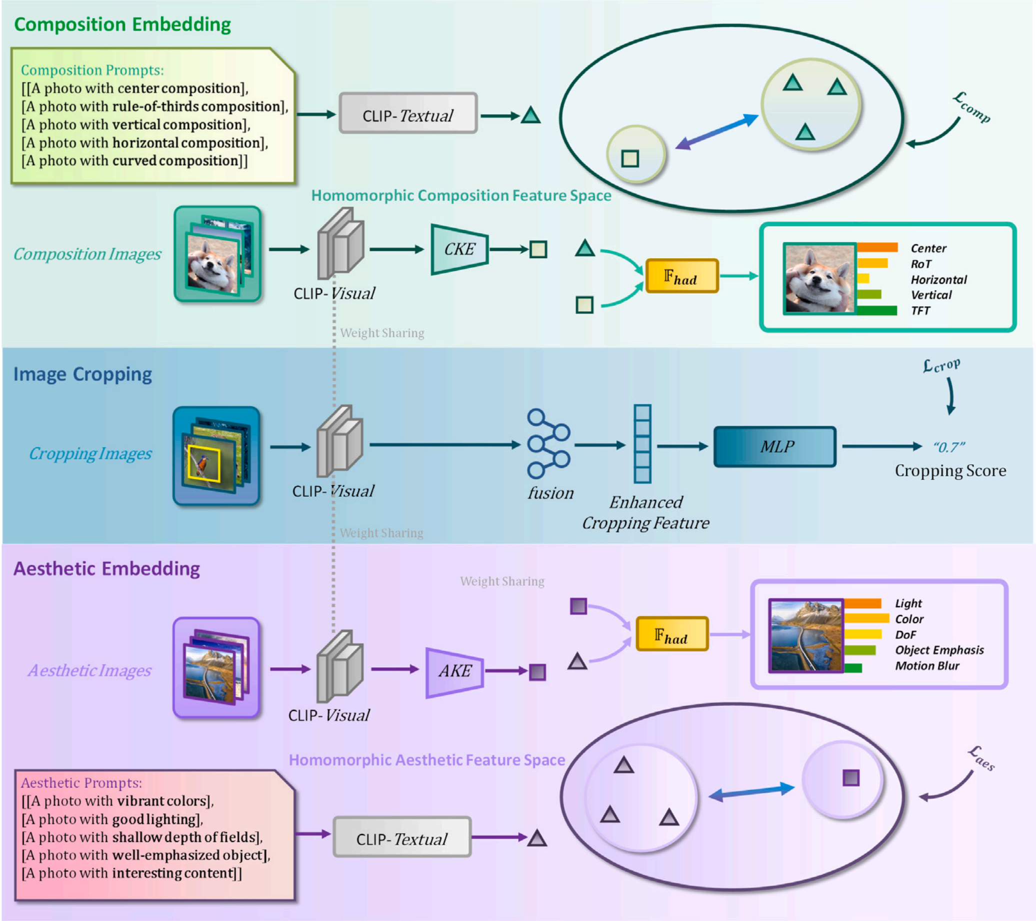

Composition Embedding
Composition Prompts:
[[A photo with center composition],
[A photo with rule-of-thirds composition],
[A photo with vertical composition],
[A photo with horizontal composition],
[A photo with curved composition]]
CLIP-Textual
$\mathcal{L}_{comp}$
Homomorphic Composition Feature Space
Composition Images
CLIP-Visual
CKE
$\mathbb{F}_{had}$
Center
RoT
Horizontal
Vertical
TFT
Weight Sharing
Image Cropping
$\mathcal{L}_{crop}$
Cropping Images
CLIP-Visual
fusion
Enhanced Cropping Feature
MLP
"0.7"
Cropping Score
Weight Sharing
Aesthetic Embedding
Weight Sharing
Aesthetic Images
CLIP-Visual
AKE
$\mathbb{F}_{had}$
Light
Color
DoF
Object Emphasis
Motion Blur
Homomorphic Aesthetic Feature Space
$\mathcal{L}_{aes}$
Aesthetic Prompts:
[[A photo with vibrant colors],
[A photo with good lighting],
[A photo with shallow depth of fields],
[A photo with well-emphasized object],
[A photo with interesting content]]
CLIP-Textual

Fig. 2. Overview of CLIPCropping framework. There are three branches in CLIPCropping, namely **(a)** Composition Embedding Learning, **(b)** Aesthetic Embedding Learning, and **(c)** Image cropping. During the training process, CLIP, CKE and AKE are optimized through $\mathcal{L}_{comp}$ and $\mathcal{L}_{aes}$. Then the CLIP and MLP are optimized through $\mathcal{L}_{crop}$. Note that Only **(c)** Image Cropping is used during the inference process.

knowledge, provides labels ranging from −1 to 1, images with negative labels for the five compositional attributes are considered as negative samples with respect to the corresponding textual prompt, while those with positive labels are treated as positive sample pairs.

In optimizing $\mathbb{F}_{\text{CLIP-visual}}$ and $\mathbb{F}_{\text{AKE}}$, we use a subset of the AADB [62] dataset $\mathcal{D}_{aes} = \{\boldsymbol{x}^i_{aes}, \boldsymbol{y}^i_{aes}\}^{N_{aes}}_{i=1}$, which provides the aesthetic images $\boldsymbol{x}_{aes}$ and their associated five aesthetic attributes $\boldsymbol{y}_{aes}$, including vibrant color, good lighting, shallow depth of field, well-emphasized object, and interesting content. The encoding process is carried out contrastively, similar to that of CKE:

$$\mathcal{L}_{\text{aes}}(\phi^{visual}_{\text{aes}}, \phi^{textual}_{\text{aes}}; \Theta_{\text{aes}}) = -\log \frac{\exp(\text{sim}(\phi^{visual}_{\text{aes}}, \phi^{textual+}_{\text{aes}}))}{\sum_{j=0}^{5} \exp(\text{sim}(\phi^{visual}_{\text{aes}}, \phi^{textual}_{\text{aes},j}))}, \tag{6}$$

where $\Theta_{\text{aes}} = \{\theta_{\text{AKE}}, \theta_{\text{CLIP}}, \theta^{aes}_{\text{learner}}\}$, $\phi^{textual}_{\text{aes}}, \phi^{textual+}_{\text{aes}}, \phi^{visual}_{\text{aes}}$ is the textual aesthetic prompt feature, the ground truth aesthetic prompt feature, and aesthetic visual feature, respectively. $\phi^{visual}_{\text{aes}}$ is obtained through $\mathbb{F}_{\text{CLIP-visual}}$ and $\mathbb{F}_{\text{AKE}}$ with a aesthetic image $\boldsymbol{x}_{\text{aes}}$:

$$\phi^{\text{visual}}_{\text{aes}} = \mathbb{F}_{\text{AKE}}(\mathbb{F}_{\text{CLIP-Visual}}(\boldsymbol{x}_{\text{aes}})). \tag{7}$$

and aesthetic textual feature $\phi^{textual}_{\text{aes}}$ is obtained through textual encoder $\mathbb{F}_{\text{textEnc}}$:

$$\phi^{textual}_{\text{aes}} = \mathbb{F}_{\text{texEnc}}(\mathbf{P}^{aes}_{\beta}, \sigma^{aes}_{embed}), \tag{8}$$

where learned aesthetic prompts $\mathbf{P}^{aes}_{\beta}$ is obtained through CLIP tokenizer $\mathbb{F}_{\text{CLIP-tokenizer}}$:

$$\mathbf{P}^{aes}_{\beta} = \mathbb{F}_{\text{CLIP-tokenizer}}(\mathbf{P}^{aes}_{\alpha}), \tag{9}$$

and composition visual embedding σ^{comp}_{embed} is obtained through the aesthetic prompt learner $\mathbb{F}_{learner A}$ with the composition visual feature $\phi^{\text{visual}}_{\text{aes}}$:

$$\sigma^{aes}_{embed} = \mathbb{F}^{aes}_{learner}(\phi^{\text{visual}}_{\text{aes}}). \tag{10}$$

3.3. Image cropping

The structure of the Image Cropping branch is also shown in Fig. 2. This branch accepts an image to be cropped and its candidate crops, outputs the cropping evaluation scores. In particular, for an image with corresponding candidate crops $\boldsymbol{X}_{crop} = \{(\boldsymbol{x}_{crop}, y^i_{crop})|y^i_{crop} \in C\}$, its composition aesthetic embedded cropping feature $\boldsymbol{h}_g$ is first obtained by employing the $\mathbb{F}_{\text{CLIP-Visual}}$:

$$\boldsymbol{h}_g = \mathbb{F}_{\text{CLIP-Visual}}(\boldsymbol{x}_{crop}). \tag{11}$$

Then, fusion module $\mathbb{F}_{\text{fusion}}$ map $\boldsymbol{h}_g$ to the enhanced cropping feature $\boldsymbol{h}_{\text{enh}}$:

$$\boldsymbol{h}_{\text{enh}} = \mathbb{F}_{\text{fusion}}(\boldsymbol{h}_g), \tag{12}$$

Similarly adhering to the aforementioned design philosophy, the structure of $\mathbb{F}_{\text{fusion}}$ is designed as one convolutional layer followed by RoIAlign and RoDAlign [39]. Ultimately, an MLP $\mathbb{F}_{\text{MLP}}$ maps the enhanced cropping feature $\boldsymbol{h}_{\text{enh}}$ into a cropping score $\hat{y}$:

$$\hat{y} = \mathbb{F}_{\text{MLP}}(\boldsymbol{h}_{\text{enh}}). \tag{13}$$

For optimizing $\mathbb{F}_{\text{MLP}}$, ranking loss and Huber loss were used :

$$\mathcal{L}_{crop} = \mathcal{L}_{\text{huber}} + \mathcal{L}^{mlp}_{rank}, \tag{14}$$

where $\mathcal{L}_{\text{huber}}$ is used to optimize the absolute difference of cropping results, which is computed as:

$$\mathcal{L}_{\text{huber}} = \begin{cases} \frac{1}{2N}\sum_{i=1}^{N}(y^{i}_{crop} - \hat{y}_i)^2, & \text{if } |\hat{y}_i - y^{i}_{crop}| \le 0, \\ \frac{1}{N}\sum_{i=1}^{N}|y^{i}_{crop} - \hat{y}_i|, & \text{if } |\hat{y}_i - y^{i}_{crop}| > 0, \end{cases} \tag{15}$$

where N is the number of candidate crops. The overall training process of CLIPCropping is shown in Algorithm 1.

Algorithm 1 The proposed CLIPCropping model.

```
Input: Composition training set $\mathcal{D}_{\text{comp}} = \{\boldsymbol{x}_i, \boldsymbol{y}^{i}_{\text{comp}}\}_{i=1}^{N_{\text{comp}}}$, aesthetic
    training set $\mathcal{D}_{\text{aes}} = \{\boldsymbol{x}_i, \boldsymbol{y}^{i}_{\text{aes}}\}_{i=1}^{N_{\text{aes}}}$, image cropping dataset $\mathcal{D}_{\text{crop}} =$
    $\{\boldsymbol{x}_i, \boldsymbol{C}_i, \}_{i=1}^{N_{\text{crop}}}$.
Output: Predicted cropping scores $\hat{y}$.
Initialize all the parameters of CLIPCropping, including the parameters for Composition Knowledge Embedding (CKE), Aesthetic Knowledge Embedding (AKE), and the MLP cropping regressor.
\\ Composition Embedding Learning
for $i = 1, 2, \ldots, N_{\text{comp}}$ do
    Sample $x_i$ and $y_{\text{comp}}$ from $\mathcal{D}_{\text{comp}}$
    Extract visual composition feature $\phi^{\text{visual}}_{\text{comp}}$ using Eq. (2)
    Extract compositional prompt feature $\phi^{\text{textual}}_{\text{comp}}$ using Eqs. (3), (4), and (5)
    Update compositional parameters $\Theta_{\text{comp}}$ by minimizing the contrastive loss using Eq. (1)
end for
\\ Aesthetic Embedding Learning
for $i = 1, 2, \ldots, N_{\text{aes}}$ do
    Sample $x_i$ and $y_{\text{aes}}$ from $\mathcal{D}_{\text{aes}}$
    Extract visual aesthetic feature $\phi^{\text{visual}}_{\text{aes}}$ using Eq. (7)
    Extract aesthetic prompt feature $\phi^{\text{textual}}_{\text{aes}}$ using Eqs. (8), (9), and (10)
    Update aesthetic parameters $\Theta_{\text{aes}}$ by minimizing the contrastive loss using Eq. (6)
end for
\\ Aesthetic Image Cropping
for $i = 1, 2, \ldots, N_{\text{crop}}$ do
    Sample $x_i$ and $y_{\text{crop}}$ from $\mathcal{D}_{\text{crop}}$
    Extract cropping features and predict scores $\hat{y}$ using Eqs. (11), (12), and (13)
    Update cropping parameters $\Theta_{\text{crop}}$ by minimizing the loss using Eq. (14)
end for
```

3.4. Loss functions

the overall loss function $\mathcal{L}$ in CLIPCropping is a weighted combination of the losses from these branches:

$$\mathcal{L} = \mathcal{L}_{comp} + \mathcal{L}_{aes} + \mathcal{L}_{crop}. \tag{16}$$

4. Experimental results

An experimental analysis of CLIPCropping is presented in the next section, which includes a series of evaluations and studies to demonstrate its effectiveness and robustness. We provide a detailed description of the datasets used in the evaluation, followed by an explanation of the standardized evaluation methodology employed for performance assessment. Subsequently, we conduct a comprehensive performance evaluation and compare CLIPCropping with state-of-the-art methods available. Finally, an ablation study is conducted to dissect the contributions of various components in our approach, providing insights into their impact on overall performance and underlying mechanisms.

4.1. Datasets

GAICD-1236 [63]. This dataset contains 1,236 images, of which 1,000 were cropped to improve composition significantly. Nineteen experienced photographers and art school students were involved in the labeling process. The dataset contains 106,860 candidate cropping frames, each of which was labeled by seven annotators according to the criteria of "bad", "poor", "fair", "good", and "excellent". The mean opinion score (MOS) of these seven annotators is the final actual value (GT). Each image contains around 90 candidate crops with cropping scores. Following the official protocol, we used 1,036 images for training and 200 images for testing.

GAICD-3336 [39]. This dataset contains 3,336 images, 3,000 of which can be cropped to improve the composition significantly. A total of fifteen annotations labeled these images, and they were either undergraduate students or graduate students majoring in art. All annotators provided each candidate crops with "bad", "poor", "fair", "good", and "excellent" scores, and GT is the MOS of these annotators. We follow the official protocol during the training and test phases.

FCDB [33]. This dataset consists of 1,536 images. Unlike the two aforementioned AIC datasets, FCDB includes one optimal crop and ten ranked pairs for each image. Following the official protocol, we utilized 1,225 images for training and 311 for testing. The training phase is similar with above two evaluation-based datasets. During the testing phase, CLIPCropping takes 21 crop candidates (the optimal crop and 10 ranked pairs) as input, predicts the cropping scores of these candidates, and then selects the one with the highest score as the best crop.

CADB [59]. This dataset contains several composition labels. We select five composition elements, including center, rule of thirds, horizontal, vertical, and curved, which can characterize the differences before and after cropping. Finally, we used 10,429 images for training and 1,043 for testing.

AADB [62]. This dataset contains aesthetic labels. We also select five aesthetic elements, including color, lighting, depth of field, well-emphasized object, and interesting content, which can characterize the differences before and after cropping. Finally, we used 8,458 images for training and 1,000 for testing.

4.2. Evaluation protocols

To evaluate the alignment between the predicted scores and the ground truth Mean Opinion Scores (MOS), we employ the Pearson Linear Correlation Coefficient (PLCC) and Spearman's Rank-Order Correlation Coefficient (SRCC). The PLCC is used to measure the accuracy of the predictions, indicating how closely the predicted values match the actual MOS values on a linear scale. In contrast, the SRCC assesses the monotonic consistency of the predictions, reflecting how well the rank order of the predicted scores preserves the rank order of the ground truth MOS. These metrics serve as benchmarks for assessing the performance of our predictive models, with PLCC focusing on accuracy and SRCC on the ordinal relationship between predicted and actual scores.

The $\text{Acc}_{K/N}$ proposed by Zeng et al. [39], is a straightforward metric that measures the percentage of images correctly cropped by the model. $\text{Acc}_{K/N}$ can be computed through:

$$\text{Acc}_{K/N} = \frac{1}{TK}\sum_{i=1}^{T}\sum_{k=1}^{K}\mathbb{1}(c_{i,k} \in \mathbb{S}_i(N)), \tag{17}$$

where $\{c_{i,k}\}_{k=1}^{K}$ denote all the K candidate crops of image i. $\mathbb{1}(\cdot) = 1$ if the boolean expression is true, otherwise $\mathbb{1}(\cdot) = 0$.

The metric $\mathrm{Acc}^{w}_{K/N}$ is an effective complement for metric $\mathrm{Acc}_{K/N}$, which can distinguish the ranking among the returned Top-N crops. This metric is defined as:

$$\mathrm{Acc}^{w}_{K/N} = \frac{1}{TK}\sum_{i=1}^{T}\sum_{k=1}^{K}\mathbb{1}(c_{i,k} \in \mathbb{S}_i(N)) * \omega_{i,j}, \tag{18}$$

where

$$\omega_{i,j} = e^{\frac{-(v_{i,j}-j)}{N}}. \tag{19}$$

When the dataset contains only one best crop, following prior art [45], mean Intersection over Union (mIOU) and mean Displacement (mDisp) are used for evaluation. Both metrics assess the difference between the best crop and the Ground Truth (GT). The difference is that mIOU focuses on the overlap between the area of the predicted best crop and the area of the GT, quantifying the accuracy of the spatial alignment. In contrast, mDisp focuses on the difference between the absolute values of the coordinates, measuring the precise positional deviation of the predicted best crop from the GT. Given N predicted crops $\hat{y}_i, i = 1, \ldots, N$ with coordinates $\{\hat{b^i_1}, \hat{b^i_2}, \hat{b^i_3}, \hat{b^i_4}\}, i = 1, \ldots, N$ and corresponding GT $y_i = \{b^i_1, b^i_2, b^i_3, b^i_4\}, i = 1, \ldots, N$, mIoU can be calculated as :

$$\mathrm{mIoU} = \frac{1}{N}\sum_{i}\mathrm{IoU}_i, \tag{20}$$

where IoU_i can be calculated as :

$$\mathrm{IoU}_i = \frac{M^p \bigcap M^{gt}}{M^p \bigcup M^{gt}}, \tag{21}$$

where M^p and M^{gt} denote the area of the predicted and GT candidate crop respectively.

mDisp can be computed as :

$$\mathrm{mDisp} = \frac{1}{N}\sum_{i}\mathrm{Disp}_i, \tag{22}$$

where Disp_i can be calculated as :

$$\mathrm{Disp}_i = \frac{1}{4}\sum_{j}\|b_j - \hat{b}_j\|_1. \tag{23}$$

In evaluating all candidate crops, the PLCC is crucial for measuring accuracy, while the SRCC is essential for assessing the consistency of the ranking order. Conversely, the $\mathrm{Acc}_{K/N}$ and $\mathrm{Acc}^{w}_{K/N}$ metrics primarily evaluate the model's efficacy in predicting the Top-K selections, ensuring the most satisfactory crops are identified. Elevated absolute values higher PLCC, SRCC, $\mathrm{Acc}_{K/N}$ and $\mathrm{Acc}^{w}_{K/N}$ are indicative of superior model performance. To align existing results, PLCC, SRCC, and Acc series are used in GAICD-1236, and GAICD-3336. mIOU and mDisp are used in FCDB.

4.3. Implementation details

CLIPCropping is implemented using PyTorch 2.0.1, Python 3.10. We train CLIPCropping with a 12th Gen Intel(R) Core(TM) i7-12700F CPU and an NVIDIA GeForce RTX 3090 GPU. For the backbone of our model, we employ the ResNet-50 (RN50) pre-trained by OpenAI. The network is optimized using the AdamW optimizer with the following hyperparameters: learning rate of 1e−7 (Composition Branch), 1e−6 (Aesthetic Branch), and 5e−6 (Cropping Branch). Training Epoch of 10 (Composition Branch), 10 (Aesthetic Branch), and 60 (Cropping Branch). Additionally, we do not resize the short side of the source image to 256 while keeping the aspect ratio, as the CLIP-RN50 requires a fixed-size input of 224 × 224. Therefore, all input images are resized to 224 × 224 during preprocessing. Conventional data augmentation can disrupt the semantics related to the image cropping task. Therefore, during training, we do not apply operations like flipping or center cropping.

4.4. Performance evaluation

To assess the efficacy of CLIPCropping, we first discuss its quantitative metrics. Tables 1, 2, and 3, and report the experimental results, which provide structured comparisons of its performance relative to existing models. In these comparisons, the highest-performing results are shown in red, and the second-best results are shown in blue. In this competitive analysis, CLIPCropping outperforms sixteen other models.

The datasets currently used to evaluate AIC, namely GAICD-1236, GAICD-3336, and FCDB, are predominantly vision-centric, focusing on tasks in which the main change elements are spatial location and color. This can be attributed to the typical AIC task setup, usually characterized by a single visually dominant object.

For the GAICD-1236 dataset, a saliency-aware approach that utilized potential region pairings earned the second-place position. At the same time, the graph structure of the CGS model outlined the interrelationships among the possible crops, allowing it to earn the third place position. For the GAICD-3336 dataset, GAICD-v2 utilized RoIAlign, RoDAlign, and multiscale features, CGICAANet combined multimodal fusion to refine the compositional features, and SFRC utilized spatial features to achieve more consistent rankings, all of which improved performance.

CLIPCropping's innovative strategy lies in integrating two visual language knowledge learning branch: Aesthetic Embedding and the Composition Embedding. These branches are adept at encapsulating aesthetic and compositional representations, enabling CLIPCropping to surpass the capabilities of CNN-based alternatives in the image cropping domain.

4.5. Visualization results

Some of the visualization results of CLIPCropping are presented in Fig. 3. The changes in aesthetic and compositional elements before and after cropping are obtained using the following method. The composition changes are calculated as follows, and the aesthetic changes are computed similarly. For the visual composition feature $\boldsymbol{h}^{vis}_{comp}$ of the cropped image, and the corresponding normalized compositional text feature $\boldsymbol{h}^{tex}_{comp}$, the scores of the compositional elements are calculated using the Hadamard product:

$$s_{comp} = \mathbb{F}_{\mathrm{had}}(\phi^{vis}_{comp}, \phi^{tex}_{comp}), \tag{24}$$

where $\mathbb{F}_{\mathrm{had}}$ is the Hadamard product of the given vectors. Similarly, the scores for aesthetic elements $\boldsymbol{h}^{vis}_{aes}$ can be obtained using the Hadamard product. It is important to note that only the relative magnitude of these scores is meaningful.

We selected the three elements with the most significant changes for each image. The results demonstrate that CLIPCropping can adapt to a variety of image types, such as nature, portraits of people, and animals, providing high-quality cropping results. For images with large white spaces, CLIPCropping tends to enhance the central composition, making the subject more prominent while also optimizing color characteristics. For animal images, CLIPCropping adjusts the composition based on the relative position of the subject, either optimizing the central composition or applying the rule of thirds as needed. For natural scene images, where there may be no single obvious subject, CLIPCropping focuses on achieving a visually balanced composition by enhancing elements like color and lighting.

4.6. Ablation studies

To provide a clearer understanding of the necessity and motivation behind AKE and CKE, we conducted ablation studies by removing each module independently. As shown in Table 4, the full model, which incorporates both AKE and CKE, achieves the best overall performance. When AKE is removed, the model shows a significant drop in aesthetic quality, indicating that subjective aspects such as color, lighting, and

Table 1
Quantitative comparison between different methods on GAICD-1236[63]. "–" means that the original paper's results are unavailable. The best performance results are marked in **bold** and the second performance results are marked with underline.

Method	SRCC	$\text{Acc}_{1/5}$	$\text{Acc}_{2/5}$	$\text{Acc}_{3/5}$	$\text{Acc}_{4/5}$	$\overline{Acc_5}$	$\text{Acc}_{1/10}$	$\text{Acc}_{2/10}$	$\text{Acc}_{3/10}$	$\text{Acc}_{4/10}$	$\overline{Acc_{10}}$
Baseline (CVPR'19)	–	24.5	–	–	–	–	41.0	–	–	–	–
VFN (MM'17) [64]	0.450	27.0	30.0	26.0	17.5	25.1	39.0	40.5	39.0	31.5	37.5
A2-RL (CVPR'18) [43]	–	23.0	–	–	–	–	38.5	–	–	–	–
VPN (CVPR'18) [34]	–	40.0	–	–	–	–	49.5	–	–	–	–
VEN (CVPR'18) [34]	0.621	40.5	37.5	38.5	36.5	38.1	54.0	51.5	50.5	47.0	50.8
GAIC-v1 (CVPR'19) [63]	0.735	53.5	47.0	44.5	41.5	46.6	71.5	66.0	66.5	58.0	65.5
ASM-Net (AAAI'20)[36]	–	54.3	–	–	–	–	71.5	–	–	–	–
CGS(CVPR'20) [35]	–	63.0	–	–	–	–	81.5	–	–	–	–
MFDM (ESA'21) [65]	–	66.0	–	–	–	–	83.0	–	–	–	–
CLIPCropping (Ours)	**0.866**	**70.0**	**66.7**	**63.0**	**60.0**	**64.9**	**87.5**	**83.7**	**80.5**	**78.5**	**82.5**

Table 2
Quantitative comparison between different methods on FCDB [33]. The best performance results are marked in **bold** and the second performance results are marked with underline.

Method	Publication	mDisp	mIoU
VFN [64]	MM'17	0.084	0.685
DIC [66]	TPAMI'18	0.080	0.660
VEN [34]	CVPR'18	0.072	0.735
VPN [34]	CVPR'18	0.073	0.711
A2-RL [43]	CVPR'18	0.089	0.664
CGS [35]	CVPR'20	0.079	0.685
ASM [36]	AAAI'20	0.068	0.749
CACNet [45]	CVPR'21	0.069	0.718
TransView [37]	ICCV'21	0.080	0.682
GAIC-v2 [39]	TPAMI'22	0.084	0.672
SFRC [42]	CVPR'23	0.075	0.695
C2C [67]	AAAI'23	0.069	0.718
CLIPCropping (Ours)	–	**0.061**	**0.764**

Table 3
Quantitative comparison between different methods on GAICD-3336[39]. "–" means that the original paper's results are unavailable. The best performance results are marked in **bold** and the second performance results are marked with underline.

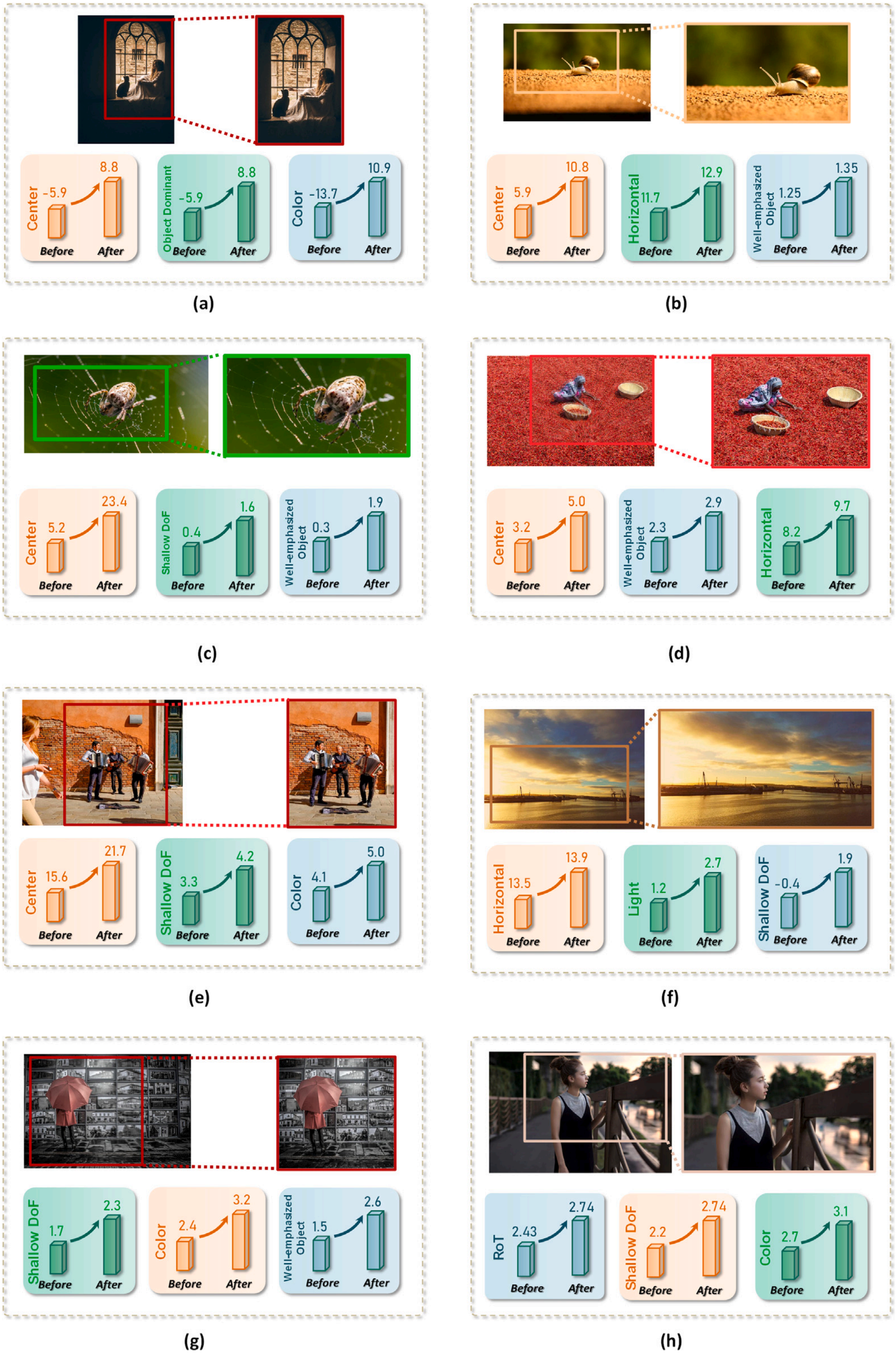
Methods	Baseline	A2-RL [43]	VPN [34]	VFN	VEN [34]	GAIC-v1[63]	CGS [35]	TransView [37]	GAIC-v2 [39]	CGICAANet [40]	SFRC [42]	Ours
Source	–	CVPR'18	CVPR'18	MM'17	CVPR'18	CVPR'19	CVPR'20	ICCV'21	TPAMI'22	TMM'23	CVPR'23	–
$\text{Acc}_{1/5}$	26.5	23.2	36.0	26.6	37.5	65.8	67.5	69.0	68.0	68.4	70.0	**72.0**
$\text{Acc}_{2/5}$	–	–	–	26.5	35.0	61.4	63.2	66.9	64.1	65.4	66.9	**70.1**
$\text{Acc}_{3/5}$	–	–	–	26.7	35.3	57.6	60.1	61.9	60.7	62.1	62.5	**66.5**
$\text{Acc}_{4/5}$	–	–	–	25.7	34.3	54.4	57.0	57.8	56.6	58.3	59.8	**63.4**
$\overline{Acc_5}$				26.4	35.5	59.8	62.0	63.9	62.4	63.6	64.8	**68.0**
$\text{Acc}_{1/10}$	44.0	39.5	48.5	40.6	50.5	85.8	83.0	85.4	85.8	84.4	86.8	**87.2**
$\text{Acc}_{2/10}$	–	–	–	40.2	49.2	82.5	80.3	84.1	82.5	83.3	84.5	**87.4**
$\text{Acc}_{3/10}$	–	–	–	40.3	48.4	80.5	78.3	81.3	80.5	81.2	82.9	**85.2**
$\text{Acc}_{4/10}$	–	–	–	39.3	46.4	77.8	76.7	78.6	77.8	78.8	79.8	**83.5**
$\overline{Acc_{10}}$	–	–	–	40.1	48.6	81.65	79.6	–	81.7	81.9	83.3	**85.8**
$\text{Acc}^{w}_{1/5}$	16.4	15.1	19.1	18.0	20.2	49,2	49.4	–	49.2	51.3	–	**51.7**
$\text{Acc}^{w}_{1/5}$	–	–	–	13.1	15.2	47.6	47.0	–	47.6	49.6	–	**51.1**
$\text{Acc}^{w}_{1/5}$	–	–	–	12.3	14.1	46.0	45.1	–	46.0	47.8	–	**50.3**
$\text{Acc}^{w}_{1/5}$	–	–	–	11.3	13.4	43.2	43.5	–	43.2	45.4	–	**48.8**
$\overline{Acc^{w}_{5}}$	–	–	–	13.7	15.7	46.5	46.2	–	46.5	48.5	–	**50.4**
$\text{Acc}^{w}_{1/10}$	27.7	25.6	29.4	27.9	30.1	65.1	64.3	–	65.1	66.8	–	**67.8**
$\text{Acc}^{w}_{2/10}$	–	–	–	22.9	25.4	64.5	62.9	–	64.5	65.9	–	**68.3**
$\text{Acc}^{w}_{3/10}$	–	–	–	21.8	24.1	63.4	61.7	–	63.4	64.1	–	**67.9**
$\text{Acc}^{w}_{4/10}$	–	–	–	20.6	23.3	61.4	60.6	–	61.4	62.5	–	**67.0**
$\overline{Acc^{w}_{10}}$	–	–	–	23.3	25.77	63.6	62.4	–	63.6	64.8	–	**67.7**
SRCC	–	–	–	0.485	0.616	0.850	0.854	0.857	0.850	0.855	0.872	**0.910**
PLCC	–	–	–	0.503	0.662	0.872	0.879	0.880	0.872	0.876	0.893	**0.920**

blur, which AKE captures, are essential for producing visually pleasing results. On the other hand, removing CKE leads to a noticeable decline in compositional quality, as the model fails to adhere to objective rules such as the center, horizontal, and vertical, which CKE is designed to capture.

The motivation for including both modules is to ensure that the model accounts for both subjective and objective aspects of image evaluation. AKE focuses on the subjective, perceptual elements of beauty, while CKE ensures adherence to established compositional guidelines. The ablation results demonstrate that using one module without the

Table 4
Ablation studies on GAICD-1236 database.

CKE	AKE	$Acc_{1/5}$	$Acc_{2/5}$	$Acc_{3/5}$	$Acc_{4/5}$	$\overline{Acc_5}$	$Acc_{1/10}$	$Acc_{2/10}$	$Acc_{3/10}$	$Acc_{4/10}$	$\overline{Acc_{10}}$	SRCC	PLCC
✗	✓	24.5	23.7	21.3	22.4	23.0	38.0	37.0	35.5	36.4	36.7	0.501	0.525
✓	✗	38.5	37.0	35.8	35.2	36.6	56.0	53.5	52.8	51.5	53.5	0.686	0.721
✓	✓	63.5	55.2	52.7	50.7	55.5	75.5	72.3	70.3	69.6	71.9	0.775	0.797



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 3. Visualization of results provided by proposed CLIPCropping. For five aesthetic elements (i.e., light, color, depth of field, object emphasis, and blur) and five composition elements (i.e., center, rule of thirds, fill the frame, horizontal, and vertical). We choose Top-3 most changing elements for visualization. All the results are from test set of GAICD-1236 [63], GAICD-3336 [39], and FCDB [33].

other leads to suboptimal results: the model without AKE lacks aesthetic appeal, and without CKE, it lacks structural integrity. Therefore, both modules are crucial for generating high-quality, well-balanced cropping results that are both aesthetically pleasing and compositionally sound.

5. Limitations and future work

Despite the strong performance of CLIPCropping shown in our experiments, there are several limitations in the current study. First, the datasets used for evaluation primarily consist of images that follow common composition rules and aesthetic principles, which may not fully represent the diversity of real-world images. For instance, images with more complex structures or abstract content might pose new challenges to the model's ability to evaluate aesthetics. Furthermore, the current version of CLIPCropping assumes a fixed input size of 224 × 224, which may limit its performance on high-resolution images where fine details are important.

In terms of future work, incorporating multi-scale processing or dynamic aspect ratio handling could further improve the model's performance on images with varying resolutions and compositions. Another promising direction would be to explore the integration of user-specific aesthetic preferences, allowing CLIPCropping to generate more personalized and adaptive cropping results.

6. Conclusion

In this paper, we have presented a new approach for the AIC task, dubbed CLIPCropping. CLIPCropping introduces the application of VLP in AIC. Our approach aligns the human decision-making process in image cropping. By combining aesthetic knowledge embedding and compositional knowledge embedding, CLIPCropping advances the field of image cropping. It improves the composition and aesthetics of images. Through extensive testing on three benchmark datasets, CLIPCropping outperforms current SOTA methods. Experiments have shown that VLP techniques can significantly improve the interpretability and validity of AIC tasks.

CRediT authorship contribution statement

Quan Yuan: Writing – original draft. **Leida Li:** Writing – review & editing. **Pengfei Chen:** Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Leida Li reports financial support was provided by Xidian University. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] Y. Huang, Q. Yuan, X. Sheng, Z. Yang, H. Wu, P. Chen, Y. Yang, L. Li, W. Lin, AesBench: An expert benchmark for multimodal large language models on image aesthetics perception, 2024, arXiv preprint arXiv:2401.08276.

[2] L. Li, Y. Huang, J. Wu, Y. Yang, Y. Li, Y. Guo, G. Shi, Theme-aware visual attribute reasoning for image aesthetics assessment, IEEE Trans. Circuits Syst. Video Technol. (TCSVT) 33 (9) (2023) 4798–4811.

[3] Y. Huang, L. Li, P. Chen, J. Wu, Y. Yang, Y. Li, G. Shi, Coarse-to-fine image aesthetics assessment with dynamic attribute selection, IEEE Trans. Multimedia (TMM) (2024) 1–14.

[4] Y. Wang, Y. Zhou, M. Li, Y. Sun, J. Ding, Blind omnidirectional image quality assessment based on semantic information replenishment, J. Vis. Commun. Image Represent. (JVCI) 103 (2024) 104241.

[5] Y. Huang, L. Li, P. Chen, J. Wu, Y. Yang, Y. Li, G. Shi, Coarse-to-fine image aesthetics assessment with dynamic attribute selection, IEEE Trans. Circuits Syst. Video Technol. (TCSVT) (2024) 1–14.

[6] Y. Zhou, Y. Ding, Y. Sun, L. Li, J. Wu, X. Gao, Perceptual information completion-based siamese omnidirectional image quality assessment network, IEEE Trans. Instrum. Meas. (TIM) 73 (2024).

[7] Y. Zhou, W. Gong, Y. Sun, L. Li, K. Gu, J. Wu, Quality assessment for stitched panoramic images via patch registration and bidimensional feature aggregation, IEEE Trans. Multimedia (TMM) 26 (2024) 354–3365.

[8] Y. Zhou, L. Li, S. Wang, J. Wu, Y. Fang, X. Gao, No-reference quality assessment for view synthesis using dog-based edge statistics and texture naturalness, IEEE Trans. Image Process. (TIP) 28 (9) (2019) 4566–4579.

[9] Y. Zhou, W. Gong, Y. Sun, L. Li, J. Wu, X. Gao, Pyramid feature aggregation for hierarchical quality prediction of stitched panoramic images, IEEE Trans. Multimedia (TMM) 25 (2023) 4177–4186.

[10] Y. Zhou, W. Gong, Y. Sun, L. Li, J. Wu, X. Gao, Photo aesthetics ranking network with attributes and content adaptation, IEEE Trans. Multimedia (TMM) 25 (2023) 4177–4186.

[11] Y. Zhou, Y. Sun, L. Li, K. Gu, Omnidirectional image quality assessment by distortion discrimination assisted multi-stream network, IEEE Trans. Circuits Syst. Video Technol. (TCSVT) 32 (4) (2022) 1767–1777.

[12] M. Gao, Q. Dong, Adaptive conditional denoising diffusion model with hybrid affinity regularizer for generalized zero-shot learning, IEEE Trans. Circuits Syst. Video Technol. (TCSVT) (2024) 1.

[13] T. Wang, K. Zhang, T. Shen, W. Luo, B. Stenger, T. Lu, Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method, in: Proc. AAAI Conf. Artif. Intell., AAAI, Vol. 37, (3) 2023, pp. 2654–2662.

[14] K. Wang, M. Gharbi, H. Zhang, Z. Xia, E. Shechtman, Semi-supervised parametric real-world image harmonization, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR, 2023, pp. 5927–5936.

[15] X. Yang, F. Lv, F. Liu, G. Lin, Self-training vision language BERTs with a unified conditional model, IEEE Trans. Circuits Syst. Video Technol. (TCSVT) 33 (8) (2023) 3560–3569.

[16] X. Wang, F. Shao, Q. Jiang, X. Chai, X. Meng, Y. Ho, List-wise rank learning for stereoscopic image retargeting quality assessment, IEEE Trans. Multimedia (TMM) 24 (2022) 1595–1608.

[17] Z. Peng, Q. Jiang, F. Shao, W. Gao, W. Lin, Lggd+: Image retargeting quality assessment by measuring local and global geometric distortions, IEEE Trans. Circuits Syst. Video Technol. (TCSVT) 32 (6) (2022) 3422–3437.

[18] D. Song, J. Zeng, M. Liu, X. Li, A. Liu, Fashion customization: Image generation based on editing clue, IEEE Trans. Circuits Syst. Video Technol. (TCSVT) 34 (6) (2024) 4344–4444.

[19] Y. Tewel, R. Gal, G. Chechik, Y. Atzmon, Key-locked rank one editing for text-to-image personalization, in: ACM SIGGRAPH Conf. Proc. (ACM SIGGRAPH), 2023, pp. 1–11.

[20] B. Suh, H. Ling, B. Bederson, D. Jacobs, Automatic thumbnail cropping and its effectiveness, in: Proc. Annu. ACM Symp. User Interface Softw. Technol., UIST, 2003, pp. 95–104.

[21] F. Stentiford, Attention based auto image cropping, in: Int. Conf. Comput. Vis. Syst., ICVS, 2007.

[22] I. McManus, F. Zhou, S. l'Anson, L. Waterfield, K. Stöver, R. Cook, The psychometrics of photographic cropping: The influence of colour, Mean. Expertise Percept. 40 (3) (2011) 332–357.

[23] J. Ma, Y. He, F. Li, L. Han, C. You, B. Wang, Segment anything in medical images, Nat. Commun. (NC) 15 (1) (2024) 654.

[24] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Wang, L. Wang, J. Gao, Y. Lee, Segment everything everywhere all at once, Adv. Neural Inf. Process. Syst. (NeurIPS) 36 (2024).

[25] J. Koh, D. Fried, R. Salakhutdinov, Generating images with multimodal language models, Adv. Neural Inf. Process. Syst. (NeurIPS) 36 (2024).

[26] D. Li, J. Li, S. Hoi, BLIP-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing, Adv. Neural Inf. Process. Syst. (NeurIPS) 36 (2024).

[27] C. Yan, Y. Hao, L. Li, J. Yin, A. Liu, Z. Mao, Z. Chen, X. Gao, Task-adaptive attention for image captioning, IEEE Trans. Circuits Syst. Video Technol. (TCSVT) 32 (1) (2024) 43–51.

[28] L. Wang, H. Qiu, B. Qiu, F. Meng, Q. Wu, H. Li, TridentCap: Image-fact-style trident semantic framework for stylized image captioning, IEEE Trans. Circuits Syst. Video Technol. (TCSVT) 34 (5) (2023) 3563–3575.

[29] S. Yan, N. Dong, L. Zhang, J. Tang, CLIP-driven fine-grained text-image person re-identification, IEEE Trans. Image Process. (TIP) (2023).

[30] Y. Zhou, W. Gong, Y. Sun, L. Li, K. Gu, J. Wu, Bi-level deep mutual learning assisted multi-task network for occluded person re-identification, IET Image Process. (IETIP) 17 (4) (2023) 979–987.

[31] L. Guo, S. Huang, H. Liu, B. Wen, Towards robust image denoising via flow-based joint image and noise model, IEEE Trans. Circuits Syst. Video Technol. (TCSVT) (2023) 1.

[32] X. Zhu, K. Guo, S. Ren, B. Hu, M. Hu, H. Fang, Lightweight image super-resolution with expectation-maximization attention mechanism, IEEE Trans. Circuits Syst. Video Technol. (TCSVT) 32 (3) (2021) 1273–1284.

[33] Y. Chen, T. Huang, K. Chang, Y. Tsai, H. Chen, B. Chen, Quantitative analysis of automatic image cropping algorithms: A dataset and comparative study, in: IEEE Winter Conf. Appl. Comput. Vis., WACV, 2017, pp. 226–234.

[34] Z. Wei, J. Zhang, X. Shen, Z. Lin, R. Mech, M. Hoai, D. Samaras, Good view hunting: Learning photo composition from dense view pairs, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR, 2018, pp. 5437–5446.

[35] D. Li, J. Zhang, K. Huang, M. Yang, Composing good shots by exploiting mutual relations, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR, 2020, pp. 4213–4222.

[36] Y. Tu, L. Niu, W. Zhao, D. Cheng, L. Zhang, Image cropping with composition and saliency aware aesthetic score map, in: Proc. AAAI Conf. Artif. Intell., AAAI, vol. 34, 2020, pp. 12104–12111.

[37] Z. Pan, Z. Cao, K. Wang, H. Lu, W. Zhong, TransView: Inside, outside, and across the cropping view boundaries, in: Proc. IEEE/CVF Int. Conf. Comput. Vis., ICCV, 2021, pp. 4218–4227.

[38] X. Zhang, Z. Li, J. Jiang, Emotion attention-aware collaborative deep reinforcement learning for image cropping, IEEE Trans. Multimedia (TMM) 23 (2020) 2545–2560.

[39] H. Zeng, L. Li, Z. Cao, L. Zhang, Grid anchor based image cropping: A new benchmark and an efficient model, IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 44 (3) (2022) 1304–1319.

[40] S. Ni, F. Shao, X. Chai, H. Chen, Y. Ho, Composition-guided neural network for image cropping aesthetic assessment, IEEE Trans. Multimedia (TMM) 25 (2023) 6836–6851.

[41] B. Zhang, L. Zhang, A. Shai, G. Brostow, M. Cisse, G. Farinella, T. Hassner, Human-centric image cropping with partition-aware and content-preserving features, in: Proc. Eur. Conf. Comput. Vis., ECCV, 2022, pp. 181–197.

[42] C. Wang, L. Niu, B. Zhang, L. Zhang, Image cropping with spatial-aware feature and rank consistency, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR, 2023, pp. 10052–10061.

[43] D. Li, H. Wu, J. Zhang, K. Huang, A2-RL: Aesthetics aware reinforcement learning for image cropping, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR, 2018, pp. 8193–8201.

[44] P. Lu, J. Liu, X. Peng, X. Wang, Weakly supervised real-time image cropping based on aesthetic distributions, in: Proc. ACM Int. Conf. Multimedia. (ACM MM), 2020, pp. 120–128.

[45] C. Hong, S. Du, K. Xian, H. Lu, Z. Cao, W. Zhong, Composing photos like a photographer, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR, 2021, pp. 7057–7066.

[46] G. Jia, H. Huang, C. Fu, R. He, Rethinking image cropping: Exploring diverse compositions from global views, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR, 2022, pp. 2446–2455.

[47] F.-L. Chen, D.-Z. Zhang, M.-L. Han, X.-Y. Chen, J. Shi, S. Xu, B. Xu, VLP: A survey on vision-language pre-training, Mach. Intell. Res. (MIR) 20 (1) (2023) 38–56.

[48] J. Lu, D. Batra, D. Parikh, S. Lee, ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, Adv. Neural Inf. Process. Syst. (NeurIPS) 32 (2019).

[49] H. Tan, M. Bansal, LXMERT: Learning cross-modality encoder representations from transformers, in: Proc. Conf. Empir. Methods Nat. Lang. Process., EMNLP, 2019.

[50] D. Song, S. Ma, Z. Sun, S. Yang, L. Liao, KVL-BERT: Knowledge enhanced visual-and-linguistic BERT for visual commonsense reasoning, Knowl.-Based Syst. (KBS) 230 (2021) 107408.

[51] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, Learning transferable visual models from natural language supervision, in: Int. Conf. Mach. Learn., ICML, 2021, pp. 8748–8763.

[52] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: Int. Conf. Mach. Learn., ICML, 2020, pp. 1597–1607.

[53] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, J. Gao, Unified vision-language pre-training for image captioning and VQA, in: Proc. AAAI Conf. Artif. Intell., AAAI, vol. 34, (07) 2020, pp. 13041–13049.

[54] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, J. Dai, VL-BERT: Pre-training of generic visual-linguistic representations, 2019, arXiv preprint arXiv:1908.08530.

[55] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, et al., OSCAR: Object-semantics aligned pre-training for vision-language tasks, in: Proc. Eur. Conf. Comput. Vis., ECCV, 2020, pp. 121–137.

[56] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. (JMLR) 21 (1) (2020) 5485–5551.

[57] A. Zeng, M. Attarian, B. Ichter, K. Choromanski, A. Wong, S. Welker, F. Tombari, A. Purohit, M. Ryoo, V. Sindhwani, Socratic models: Composing zero-shot multimodal reasoning with language, 2022, arXiv preprint arXiv:2204.00598.

[58] W. Hoh, F. Zhang, N. Dodgson, Salient-centeredness and saliency size in computational aesthetics, ACM Trans. Appl. Percept. (TAP) 20 (2) (2023) 1–23.

[59] B. Zhang, L. Niu, L. Zhang, Image composition assessment with saliency-augmented multi-pattern pooling, in: Proc. Brit. Mach. Vis. Conf., BMVC, 2021.

[60] K. Wang, J. Yang, C. Loy, Z. Liu, Conditional prompt learning for vision-language models, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR, 2022, pp. 16816–16825.

[61] A. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, 2018, arXiv preprint arXiv:1807.03748.

[62] S. Kong, X. Shen, Z. Lin, R. Mech, C. Fowlkes, Photo aesthetics ranking network with attributes and content adaptation, in: Proc. Eur. Conf. Comput. Vis., ECCV, 2018, pp. 662–679.

[63] H. Zeng, L. Li, Z. Cao, L. Zhang, Reliable and efficient image cropping: A grid anchor based approach, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR, 2019, pp. 5949–5957.

[64] Y. Chen, J. Klopp, M. Sun, S. Chien, K. Ma, Learning to compose with professional photographs on the web, in: Proc. ACM Int. Conf. Multimedia. (ACM MM), 2017, pp. 37–45.

[65] Y. Xu, W. Xu, M. Wang, L. Li, G. Sang, P. Wei, L. Zhu, Saliency aware image cropping with latent region pair, Expert Syst. Appl. (ESA) 171 (2021) 114596.

[66] W. Wang, J. Shen, H. Ling, A deep network solution for attention and aesthetics aware photo cropping, IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 41 (7) (2018) 1531–1544.

[67] Z. Pan, Y. Chen, J. Zhang, H. Lu, Z. Cao, W. Zhong, Find beauty in the rare: Contrastive composition feature clustering for nontrivial cropping box regression, in: Proc. AAAI Conf. Artif. Intell., AAAI, 2023.