Aesthetically Relevant Image Captioning

Zhipeng Zhong[1, 3, 4, 5], **Fei Zhou**[1, 2, 3, 4, 5] **and Guoping Qiu**[1, 2, 3, 4, 5, 6]
[1]College of Electronics and Information Engineering, Shenzhen University, China
[2]Peng Cheng National Laboratory, Shenzhen, China
[3]Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen, China
[4]Shenzhen Institute for Artificial Intelligence and Robotics for Society, China
[5]Guangdong-Hong Kong Joint Laboratory for Big Data Imaging and Communication, Shenzhen, China
[6]School of Computer Science, The University of Nottingham, UK
guoping.qiu@nottingham.ac.uk

Abstract

Image aesthetic quality assessment (AQA) aims to assign numerical aesthetic ratings to images whilst image aesthetic captioning (IAC) aims to generate textual descriptions of the aesthetic aspects of images. In this paper, we study image AQA and IAC together and present a new IAC method termed Aesthetically Relevant Image Captioning (ARIC). Based on the observation that most textual comments of an image are about objects and their interactions rather than aspects of aesthetics, we first introduce the concept of Aesthetic Relevance Score (ARS) of a sentence and have developed a model to automatically label a sentence with its ARS. We then use the ARS to design the ARIC model which includes an ARS weighted IAC loss function and an ARS based diverse aesthetic caption selector (DACS). We present extensive experimental results to show the soundness of the ARS concept and the effectiveness of the ARIC model by demonstrating that texts with higher ARS's can predict the aesthetic ratings more accurately and that the new ARIC model can generate more accurate, aesthetically more relevant and more diverse image captions. Furthermore, a large new research database containing $510K$ images with over 5 million comments and $350K$ aesthetic scores, and code for implementing ARIC are available at https://github.com/PengZai/ARIC.

Introduction

Image aesthetic quality assessment (AQA) aims to automatically score the aesthetic values of images. This is very challenging because aesthetics is a highly subjective concept. AQA models are either regressors or classifiers that extract image features and output the aesthetic scores or classes (Zhao et al. 2021). Many images, especially those on photography competition websites contain both aesthetic scores and textual comments. It has been shown that including both the visual and textual information can improve AQA performances (Zhou et al. 2016; Zhang et al. 2021).

Unlike visual contents which are rather abstract, texts are much easier for human to comprehend. Image captioning, which aims to automatically generate textual descriptions of images has been extensively researched, and much progress has been achieved in recent years with the help of deep learning technology (Hossain et al. 2019). Whilst the vast majority of researchers have focused on image captions about objects and their relations and interactions, the more abstract and arguably more challenging problem of image aesthetic captioning (IAC) (Chang, Lu, and Chen 2017), which aims to generate comments about the aesthetic aspects of images, has received much less attention.

In the existing literature, image AQA and IAC are studied independent from each other. However, these are closely related areas of aesthetic visual computing, we therefore believe that jointly study them is beneficial. For example, currently IAC performances are evaluated either subjectively or based on metrics such as SPICE (Anderson et al. 2016) and BLEU (Papineni et al. 2002) which evaluate the similarity between the generated and reference (ground truth) sentences rather than the aesthetic relevance of the texts. It will be very useful if we can directly measure the aesthetic relevance of the IAC results, for example, how accurate the generated caption can predict the images aesthetic scores.

In this paper, we first contribute a large research database called DPC2022 which contains 510K images, over 5 million comments and 350K aesthetic ratings. We then present a new IAC method called Aesthetically Relevant Image Captioning (ARIC). Based on the observation that most image comments are general descriptions of image contents and not about their aesthetics, we first introduce the concept of Aesthetic Relevance Score (ARS) of a sentence. ARS consists of 5 components including scores related to aesthetic words, the length of the sentence, words describing objects, the sentiments of the sentence, and term frequency-inverse document frequency (tf-idf). A list of aesthetic and object words has been manually constructed from DPC2022. After (automatically) labelling the comments in DPC2022 with the ARS scores, we then construct an ARS predictor based on the Bidirectional Encoder Representations (BERT) language representation model (Devlin et al. 2018).

The introduction of ARS and its predictor have enabled the design of the ARIC model which includes an ARS weighted IAC loss function and an ARS based diverse aesthetic caption selector (DACS). Unlike methods in the literature that simply learned a direct mapping between the images and their comments in the database, regardless of the aesthetic relevance of the comments (in fact many of the texts have nothing to do with aesthetics), the ARIC model is constructed based on ARS weighted loss function which ensures that it learns aesthetically relevant information. Furthermore, unlike traditional methods that pick the output sentences based on the generator's confidence which is not directly based on aesthetic relevance, the introduction of the ARS has enabled the design of DACS which can output a diverse set of aesthetically highly relevant sentences. In addition, we have fine tuned the powerful pre-trained image and text matching model CLIP (Radford et al. 2021) using DPC2022 as an alternative to ARS for selecting aesthetically relevant captions.

We have performed extensive experiments. As DPC2022 is by far one of the largest AQA databases, we first provide baseline AQA results. We then present ARIC's image aesthetic captioning performances to demonstrate its effectiveness. In summary, the contributions of this paper are

1. A large image database, DPC2022, is constructed for researching image aesthetic captioning and image aesthetic quality assessment. DPC2022 is the largest dataset containing both aesthetic comments and scores.
2. A new concept, aesthetic relevance score (ARS), is introduced to measure the aesthetic relevance of sentences. Lists of key words and other statistical information for constructing the ARS model are made available.
3. Based on ARS, we have developed the new aesthetically relevant image captioning (ARIC) system capable of producing not only aesthetically relevant but also diverse image captions.

Related Work

Image aesthetic quality assessment (IAQA). One of the main challenges in AQA is the lack of large scale high quality annotated datasets. The aesthetic visual analysis (AVA) dataset (Murray, Marchesotti, and Perronnin 2012) contains $250K$ images, each has an aesthetic score, and other labels. This is still one of the most widely used databases in aesthetic quality assessment. Recent AQA systems are mostly based on deep learning neural networks and supervised learning that take the images or their textual descriptions or both as input to predict the aesthetic scores (Valenzise, Kang, and Dufaux 2022; Zhang et al. 2020).

Image aesthetic captioning (IAC). Aesthetic image captioning was first proposed in (Chang, Lu, and Chen 2017) where the authors also presented the photo critique captioning dataset (PCCD) which contains pair-wise image comment data from professional photographers. It contains 4235 images and more than sixty thousands captions. The AVA-Captions dataset (Ghosal, Rana, and Smolic 2019) was obtained by using a probabilistic caption filtering method to clean the noise of the original AVA captions. It has about $230K$ images with roughly 5 captions per image. The DPC-Captions dataset (Jin et al. 2019) contains over $150K$ images and nearly 2.5 million comments where each comment was automatically annotated with one of the 5 aesthetic attributes of the PCCD through knowledge transfer. Very recently, the Reddit Photo Critique Dataset (RPCD) was published by (Nieto, Celona, and Fernandez-Labrador 2022). This dataset contains tuples of image and photo critiques which has $74K$ images and $220K$ comments. Image aesthetic captioning is an under explored area and existing works mostly used LSTM model to generate aesthetic captions.

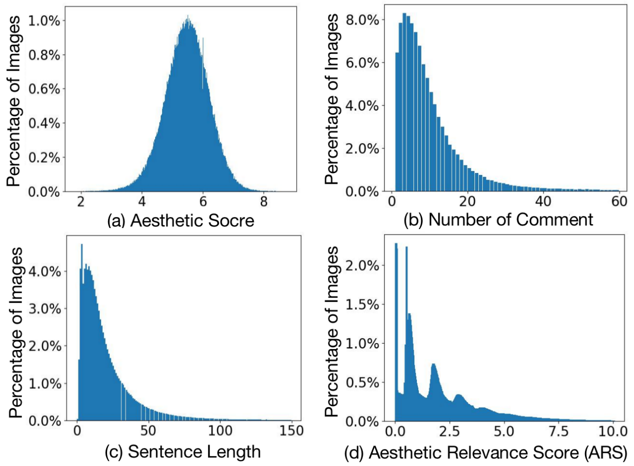
Image captioning. Image captioning aims to generate syntactically and semantically correct sentences to describe images. This is a complex and challenging task in which many deep learning-based techniques have been developed in recent years (Hossain et al. 2019). Training deep models requires large amount of annotated data which are very difficult to obtain. VisualGPT (Chen et al. 2022) leverages the linguistic knowledge from a large pre-trained language model GPT-2 (Radford et al. 2019) and quickly adapt it to the new domain of image captioning. It has been shown that this Transformer (Vaswani et al. 2017) based technique has superior performances to LSTM based methods used in previous IAC works (Chang, Lu, and Chen 2017)(Ghosal, Rana, and Smolic 2019)(Jin et al. 2019). We adopt this method for IAC to provide benchmark performances for the newly established DPC2022 dataset.

The DPC2022 Dataset

The AVA dataset (Murray, Marchesotti, and Perronnin 2012) which was constructed a decade ago remains to be one of the largest and most widely used in image AQA. As discussed above, datasets available for IAC are either small or constructed from the original AVA dataset. In the past 10 years, the source website of the AVA dataset (www.dpchallenge.com) has been continuously organising more photography competitions and has accumulated a lot more photos and comments. We therefore believe it is a good time to make full use of the data currently available from the website to construct a new dataset to advance research in image aesthetic computing. We first crawled the website and grabbed all currently available images and their comments. Initially we obtain a total of $780K$ images and their comments. We then used an industrial strength natural language processing tool *spaCy* (https://spacy.io/) to clean the data by removing items such as emoji and other strange spellings, symbols and punctuation marks. At the end, we have kept $510K$ images with good quality and clean comments. Within these $510K$ images, there are $350K$ with both comments and aesthetic scores ranging between 1 and 10. Figure 1 shows the statistics of the DPC2022 dataset. On average, each image contains roughly 10 comments, each comments on average contains 21 sentences, and the average sentence length is 19 words.

Image Aesthetic Relevance of Text

The dictionary definition of aesthetic is "concerned with beauty or the appreciation of beauty", this is a vague and subjective concept. Photographers often use visual characteristics such as lighting, composition, subject matter and colour schemes to create aesthetic photos. Comments associated with images, especially those on the internet websites, can refer to a variety of topics, not all words are relevant to the aesthetics of images, and some words may be more relevant than others. To the best of the authors' knowledge, there exist no definitions of image aesthetic relevant words and phrases in the computational aesthetic literature.



(a) Aesthetic Socre (b) Number of Comment

(c) Sentence Length (d) Aesthetic Relevance Score (ARS)

Figure 1: Basic statistics of the DPC2022 dataset.

And yet when inspecting the comments from the DPC2022 dataset, it is clear that many have nothing to do with the images' aesthetic qualities. It is therefore appropriate to distinguish words that refer to the aesthetic quality and those that are not relevant. We have developed the Aesthetic Relevant Score (ARS) to quantify a comment's aesthetic relevance. It is based on a mixture of subjective judgement and statistics from the dataset. Before describing the ARS in detail, it is appropriate to note that this is by no means the only way to define such a quantity, however, we will demonstrate its usefulness through application in image aesthetic captioning and multi-modal image aesthetic quality assessment.

Labelling a Sentence with its ARS

The *ARS* of a sentence t is defined as:

$$ARS(t) = A(t) + L(t) + O(t) + S(t) + T_{fidf}(t) \qquad (1)$$

where $A(t)$ is related to aesthetic words, $L(t)$ is related to the length of t, $O(t)$ is related to object words, $S(t)$ is the sentiment score of t, and $T_{fidf}(t)$ is related to term frequency–inverse document frequency.

The components in (1) are computed based on statistics of the DPC2022 database and details of the computational procedures are in **Appendix I** in the supplementary materials. For computing $A(t)$, we manually selected 1022 most frequently appeared image aesthetics related words such as *shot, color, composition, light, focus, background, subject, detail, contrast, etc.*, the full list of these words, $\{AW_{list}\}$, can be found in **Appendix II**. For computing $O(t)$, we manually selected 2146 words related to objects such as *eye, sky, face, ribbon, water, tree, flower, expression, hand, bird, glass, dog, hair, cat, smile, sun, window, car, etc*, the full list of these words, $\{OW_{list}\}$, can be found in **Appendix III**. The sentiment score $S(t)$ is calculated based on the $BerTweet$ model (Pérez, Giudici, and Luque 2021). Figure 1(d) shows the distribution of *ARS* of the DPC2022 dataset. It is seen that many sentences contain no aesthetic relevant information, and the majority of the comments contain very low aesthetic relevant information. Informally inspecting the data shows that this is reasonable and expected.

Automatically Predicting the ARS

For the *ARS* to be useful, we need to be able to predict any given text's aesthetic relevant score. With the labelled data described above, we adopt the pre-trained Bidirectional Encoder Representations (BERT) language representation model (Devlin et al. 2018) for this purpose. A 768 × 1 fully connected layer is cascaded to the output of BERT to predict the ARS of the input text. We train the model with the mean squared error (MSE) loss function. Table 1 shows the *ARS* prediction performance. It is seen that the Spearman's rank-order correlation (SRCC) and the Pearson linear correlation coefficient (PLCC) are both above 0.95, indicating excellent prediction accuracy. It is therefore possible to predict a sentence's ARS with this model. We call ARS predictor.

Metrics	SRCC↑	PLCC↑	RMSE↓	MAE↓
Performance	0.9553	0.9599	0.5617	0.3395

Table 1: ARS prediction accuracy.

Aesthetically Relevant Image Captioning

Image Captioning Model

For image captioning, many models based on LSTM (Vinyals et al. 2015) and Transformer (Vaswani et al. 2017) have been developed. Given an image, we first extract visual features using an encoder structure, then use a decoder to generate image captions as shown in Figure 2. To obtain image embedding, we follow the bottom-up-attention model (Anderson et al. 2018) and use ResNet-101 (He et al. 2016) as the backbone network of the Faster R-CNN (Ren et al. 2015) to extract image features. The bottom-up attention model uses a region proposal network (RPN) and a region-of-interest (ROI) pooling strategy for object detection. In this paper, we retain 50 most interesting regions and pass them to the encoder for image caption generation. Each region is represented by a 2048-dimensional feature vector.

For the decoder, we use VisualGPT (Chen et al. 2022) to generate image aesthetic captions. VisualGPT is an encoder-decoder transformer structure based on GPT-2 (Radford et al. 2019) which is a powerful language model but does not have the ability to understand images. VisualGPT uses a self-resurrecting activation unit to encode visual information into the language model, and balances the visual encoder module and the language deocder module.

Aesthetically Relevant Loss Function

Unlike general image captioning which in most cases is about revealing the content and semantics of the image, e.g., objects in the image and their relations, image aesthetic captioning (IAC) should focus on learning aesthetically relevant aspects of the image. Arguably, IAC is much more challenging than generic image captioning. The very few existing IAC related literature, e.g., (Chang, Lu, and Chen 2017), (Jin et al. 2019) and (Ghosal, Rana, and Smolic 2019), simply treat the raw image comments from the training data as the IAC ground truth. However, as discussed previously,

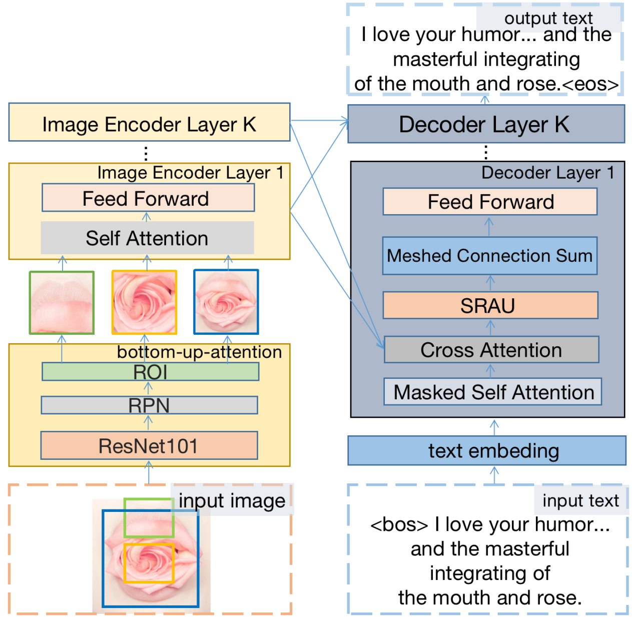


Figure 2: Image captioning model. Left: object detection is performed on the input image, visual features and positions of the regions of interest are encoded. Right: image caption is generated based on the encoded image features. Note the input text is paired with the input image for training purpose.

many of the texts contain no or very low aesthetic relevant information. The *ARS*(t) quantitatively measures the aesthetic relevance of a piece of text t, a higher *ARS*(t) indicates that t contains high aesthetic relevant information and a low *ARS*(t) indicates otherwise. We therefore use *ARS*(t) to define the aesthetically relevant loss function to construct the IAC model.

Given training set $\{T(k)\} = \{I, y^*(t_k)\}$ which contains pairs of input image I and one of its corresponding ground-truth caption sentences $y^*(t_k) = (y_1^*(t_k), y_2^*(t_k), ..., y_{N_k}^*(t_k))$ consisting of words $y_i^*(t_k)$, $i = 1, 2, ..., N_k$, we generate a caption sentence $y(t_k) = (y_1(t_k), y_2(t_k), ..., y_{N_k}(t_k))$ that maximises the following *ARS*(t) weighted cross-entropy loss

$$L_{AR}(\theta) = -\sum_{k=1}^{k=|T|} ARS(t_k) \sum_{i=1}^{N_k} \log p\left(y_i(t_k) = y_i^*(t_k)|\boldsymbol{\theta}\right) \tag{2}$$

where t_k is the k^{th} training sentence (note one image I can have multiple sentences), $|T|$ is the training set size (in terms of sentences), N_k is the number of words in the k^{th} training sentence, and $\boldsymbol{\theta}$ is the model parameters.

Diverse Aesthetic Caption Selector (DACS)

Traditional IAC models such as the one described in Figure 2 output sentences based on the generator's confidence. All existing methods in the literature (Chang, Lu, and Chen 2017), (Jin et al. 2019) and (Ghosal, Rana, and Smolic 2019) adopt this approach. However, the generator's confidence is not directly based on aesthetic relevance. The new ARS concept introduced in this paper has provided a tool to measure the aesthetic value of a sentence, which in turn can help selecting aesthetically more relevant and more diverse captions from the generator.

Based on the ARS score defined in (1) for a piece of text t, we design a sentence selector method that enables the IAC model to generate diverse and aesthetically relevant sentences. Given a picture, we use the IAC model and beam search (Olive, Christianson, and McCary 2011) to generate N most confident sentences. Then we use a sentence transformer[1] to extract the features of the sentences, and then calculate the cosine similarity among the N sentences. We group sentences whose cosine similarity is higher than 0.7 into the same group such that each group contains many similar sentences. Then, we use the *ARS* predictor to estimate the *ARS* of the sentences in each group. If the average *ARS* of a group's sentences is below the mean *ARS* of the training data (2.1787), then the group is discarded because their aesthetic values are low. From the remaining groups, we then pick the sentence with the highest *ARS* in each group as the output. With this method, which we call the Diverse Aesthetic Caption Selector (DACS), we generate aesthetically highly relevant and diverse image captions. In the experiment section, we will show that OpenAI's CLIP (Radford et al. 2021), a powerful image and text embedding model that can be used to find the text snippet best represents a given image, can be fine tuned to play the role of ARS for picking the best sentence in a group as output.

Multi-modal Aesthetic Quality Assessment

The purposes of performing multi-modal AQA are two folds. Firstly, the newly established DPC2022 is one of the largest publicly available datasets and quite unique in the sense that it contains both comments and aesthetic scores. We want to provide some AQA performance baselines. Secondly, we make the reasonable assumption that the more accurate a piece of text can predict an image's aesthetic rating, the more aesthetically relevant is the text to the image.

The image AQA model is shown in Figure 3. There are 3 separate paths. The first takes the image as input and output the image's aesthetic rating. A backbone neural network is used for feature extraction. The features are then fed to an MLP to regress the aesthete rating. The second path takes the textual description of the image as input and output the image's aesthetic score. Again, a neural network backbone is used for textual feature extraction. The features are fed to an MLP network to output the aesthetic score. The third path is used to implement multi-modal AQA where visual and textual features are concatenated and fed to a MLP to output the aesthetic score.

Experimental Results

We perform experiments based on the newly constructed DPC2022 dataset. It has a total of 510,000 photos where each photo also contains a review text. 350,000 out of the

[1] https://huggingface.co/models, **all-miniLM-L6-v2**

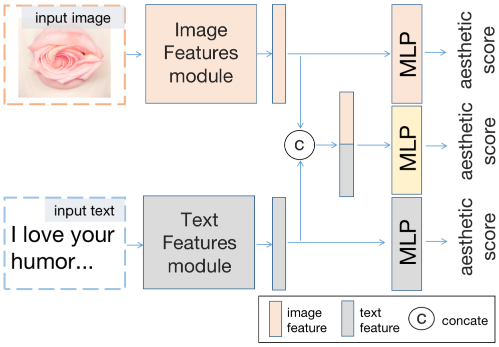


Figure 3: Multi-modal Image Aesthetic Quality Assessment.

510,000 photos also have aesthetic ratings. We call the dataset containing all images setA, and the subset containing images with both reviews and aesthetic ratings SetB. We use SetA for photo aesthetic captioning experiment and SetB for multi-modal aesthetic quality assessment. We divide SetB into a test set containing 106,971 images and a validation set containing 10,698 images, and the remaining 232,331 images are used as the training set. In photo aesthetic captioning, we use the same test set and validation set as those used in multi-modal aesthetic quality assessment, and the remaining SetA data (392,331 images) is used for training.

Evaluation Criteria

Similar to those in the literature, we use 5 metrics to measure the performance of image AQA including Binary Accuracy (ACC), Spearman rank order correlation coefficient (SRCC), Pearson linear correlation coefficient (PLCC), root mean square error (RMSE), and mean absolute error(MAE). For image captioning, we use 4 metrics widely used in related literature including CIDER (Vedantam, Lawrence Zitnick, and Parikh 2015), METETOR (Banerjee and Lavie 2005), ROUGE (Lin 2004), and SPICE (Anderson et al. 2016). SPICE is a word-based semantic similarity measure for scene graphs, and the others are all based on n-gram.

Implementation Details

For image based AQA, we have used VGG-16, RESNET18, DESNET121, RESNEXT50, and ViT (Dosovitskiy et al. 2020) as the backbone network and used their pre-trained models. For text based AQA, we have used the TEXTCNN (Kim 2014), TEXTRCNN (Lai et al. 2015), BERT (Devlin et al. 2018) and ROBERTA (Liu et al. 2019)) as the backbone network. For BERT and ROBERTA, we use their pre-trained models. Finally, we limit the number of tokens for each image's comment to 512. For IAC, we used the pre-trained GPT2 model and set token size to 64. All experiments were performed on a machine with 4 NVIDIA A100 GPUs. Adam optimizer with a learning rate of $2e^{-5}$ without weight decay was used.

Image AQA Baseline Results

In the first set of experiments, we experimented various network architectures in order to obtain some baseline image AQA results. Table 2 lists the baseline results of image based, text based, and multi-modal AQA results when the backbone used different network architectures. These results show that the textual reviews of the images are aesthetically highly relevant. In two out of 5 metrics, using the review text only gives the best performances. It is also seen that in 3 other metrics, multi-modal AQA has the best performances.

BB	ACC	SRCC	PLCC	RMSE	MAE
		Image based AQA			
VGG	0.7848	0.5914	0.6078	0.5977	0.4699
R18	0.8007	0.5946	0.6108	0.5850	0.4588
D121	0.8096	0.6331	0.6471	0.5501	0.4287
R50	0.8067	0.6386	0.6530	0.5486	0.4282
ViT	0.8194	0.6755	0.6868	0.5356	0.4193
		Text based AQA			
TCNN	0.812	0.6069	0.6251	0.5891	0.461
TR	0.8665	0.7516	0.7730	0.4649	0.3611
BE	0.8810	0.8024	0.8219	0.4235	0.3292
RO	**0.8939**	0.8334	0.8551	0.3826	0.2988
		Multi-modal (image plus text) AQA			
R50+TR	0.8591	0.8289	0.8456	0.3864	0.2998
R50+BE	0.8666	0.8493	0.8695	0.3780	0.2941
ViT+TR	0.8622	0.8386	0.8537	0.3747	0.2901
ViT+RO	0.8693	**0.8629**	**0.8803**	**0.3545**	**0.2756**
ViT+BE	0.8697	0.8594	0.8766	0.3604	0.2814

Table 2: AQA baseline results of DPC2022. VGG, R18, D121, R50, TCNN, BE, RO, TR short for VGG16, RESNET18, DESNET121, RESNEXT50, TEXTCNN, BERT, ROBERTA, TEXTRCNN, respectively.

Data Group	ACC	SRCC	PLCC	RMSE	MAE
All	0.866	0.865	0.881	0.368	0.288
Low A	0.849	0.824	0.840	0.372	0.290
High A	0.896	0.915	0.919	0.362	0.284
Low O	0.851	0.833	0.849	0.371	0.290
High O	0.895	0.90	0.912	0.363	0.284
Low L	0.852	0.823	0.837	0.370	0.289
High L	0.894	**0.918**	**0.921**	0.365	0.286
Low S	0.835	0.814	0.827	0.375	0.293
High S	**0.930**	0.901	0.914	**0.355**	**0.277**
Low T_{fidf}	0.850	0.818	0.833	0.371	0.290
High T_{fidf}	0.889	0.910	0.914	0.365	0.284
Low ARS	0.848	0.821	0.835	0.372	0.291
High ARS	0.899	0.915	0.920	0.362	0.283

Table 3: Multi-modal AQA for images with high and low aesthetic relevant comments. These results are all based on VIT+BERT model.

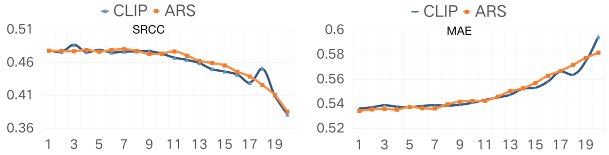
Verification of ARS

To verify the soundness of the newly introduced *ARS*, we divide the DPC2022 validation set into two groups according

ARS THR	ACC	SRCC	PLCC	RMSE	MAE
$\leq m - 1.0\sigma$	0.807	0.599	0.644	0.425	0.329
$\leq m - 0.8\sigma$	0.821	0.696	0.717	0.398	0.307
$\leq m - 0.6\sigma$	0.835	0.760	0.775	0.384	0.300
$\leq m - 0.4\sigma$	0.841	0.794	0.809	0.376	0.294
$\leq m - 0.2\sigma$	0.844	0.810	0.823	0.375	0.293
$\leq m - 0.0\sigma$	0.848	0.821	0.835	0.372	0.291
$\geq m + 0.0\sigma$	0.899	0.915	0.920	0.362	0.283
$\geq m + 0.2\sigma$	0.909	0.920	0.924	0.362	0.283
$\geq m + 0.4\sigma$	0.912	0.924	0.926	0.366	0.286
$\geq m + 0.6\sigma$	0.918	0.927	0.928	0.365	0.283
$\geq m + 0.8\sigma$	0.923	0.928	0.929	0.367	0.285
$\geq m + 1.0\sigma$	0.923	0.929	0.929	0.366	0.285
ungroup(all)	0.866	0.865	0.881	0.368	0.288

Table 4: Multi-modal image AQA. As the ARS of the comment texts increases, the aesthetic rating prediction performances increases. m_{ARS} is the mean of *ARS* and σ_{ARS} is the variance of the *ARS* of the DPC2022 dataset. $ARS \leq m_{ARS} - \alpha\sigma_{ARS}$, short for $\leq m - \alpha\sigma$, indicates the group of images each with a comment having an *ARS* value less than or equal to $m_{ARS} - \alpha\sigma_{ARS}$.

to the ARS. If the *ARS* model is sound, then we would expect a review text with a higher *ARS* should be able to predict the image's aesthetic rating more accurately. Conversely, a review text that has a lower *ARS* would produce a less accurate prediction of the image's aesthetic rating. Table 3 shows the multi-modal AQA performances for two groups of testing samples. Low ARS (A, O, L, S, T_{fidf}) represent the group where test samples have a ARS (A, O, L, S, T_{fidf}) lower than their respective average values. High ARS (A, O, L, S, T_{fidf}) represent the group where test samples have a ARS (A, O, L, S, T_{fidf}) higher than their respective average values. Table 4 shows that as the *ARS* increases, so does the AQA performances. These results clearly show that comments with high *ARS* can consistently predict the images aesthetic rating more accurately than those with low *ARS*. This means that *ARS* and its components can indeed measure the aesthetic relevance of textual comments.



(a) (b)

Figure 4: Comparison between CLIP and ARS for ranking aesthetically relevant sentences for image AQA. Horizontal-axis is the rank position.

ARS versus CLIP for Ranking Text Relevance

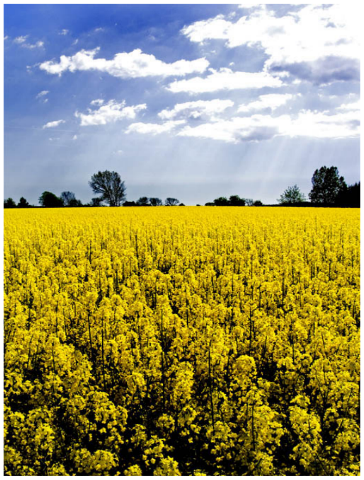
CLIP (Contrastive Language-Image Pre-Training) is a neural network trained on a variety of (image, text) pairs (Radford et al. 2021). It is a powerful model that can match natural language descriptions with visual contents. In this experiment, we first fine tune the pre-trained CLIP model with the training set of the DPC2022 data such that the images and their corresponding comments are matched. With such a fine tuned model, we can rank the sentences in the comments of the images. Let CLIP(I, t) be the matching score between an image I and a sentence t. Suppose we have two sentences t_i and t_j, if CLIP$(I, t_i) >$ CLIP(I, t_j) then t_i and I is a better match. Because CLIP has been fine tuned on DPC2022, it is reasonable to assume that if a text and an image has a better CLIP matching score, then the text can describe the image more accurately. Similar to *ARS*, we can use CLIP to rank the sentences according to their CLIP matching scores. To compare the CLIP and *ARS* in the selection of texts for image AQA, we rank the sentences of the images in the test set and perform text based AQA for the sentences in different ranks. The SRCC and MAE performances of those ranked by *ARS* and CLIP are shown in Figure 4. It is seen that the higher ranking sentences by both methods can predict the aesthetic rating more accurately, and that both seem to perform very similarly. These results show that the properly fine tuned CLIP model can also be used to select aesthetically relevant text for image AQA. Figures 5 shows visual examples of how the sentences are ranked by *ARS* and CLIP.

Method	ACC	SRCC	PLCC	RMSE	MAE
IO	0.819	0.675	0.686	0.535	0.419
AQA using Ground Truth Text					
TO	0.853	0.827	0.851	0.382	0.296
I+T	0.840	0.752	0.769	0.465	0.365
AQA using Generated Captions					
TO Trad	0.763	0.433	0.426	0.892	0.715
TO ARS	0.769	0.471	0.469	0.788	0.622
TO CLIP	0.772	0.476	0.474	0.764	0.601
I+T Trad	0.818	0.681	0.692	0.542	0.423
I+T ARS	0.819	0.683	0.693	0.535	0.419
I+T CLIP	0.820	0.683	0.694	0.534	0.417

Table 5: Image AQA using generated captions. IO, TO, I+T, Trad, ARS, CLIP short for Image only, Text only, Multi-modal(Image+Text), Traditional, DACS (ARS), DACS (CLIP) respectively. Traditional: no aesthetic relevance selection where all generated sentences are used. DACS(ARS): using sentences selected by the ARS based diverse aesthetic caption selector (DACS). DACS(CLIP): using sentences selected by the CLIP based diverse aesthetic caption selector (DACS).

Aesthetic Captioning Performances

Table 6 shows the aesthetic image captioning performances of our new aesthetically relevant model (ARIC) based on the loss function $L_{AR}(\theta)$ as defined in (2). For comparison, we have implemented a baseline model in which standard cross entropy loss function is used, i.e., setting $ARS(t_k) = 1$ in (2). It is seen that our new model consistently outperforms the baseline model. These results demonstrate the usefulness



Traditional

1. Nice shot.,
2. Great shot.,
3. Beautiful shot.,
4. Beautiful colors.,
5. Great colors.,
6. Great shot!,
7. Nice colors.,
8. Nice composition.,
9. Lovely shot.,
10. Beautiful shot!,
11. Beautiful scene.,
12. ...

ARS selector

1. Beautiful shot.
2. Beautiful colors and composition.
3. Nicely composed.
4. Beautiful scene.
5. Beautiful landscape.,
6. Beautiful sky.,
7. Looks like a painting.,
8. I love the colors and the contrast between the yellow and blue.,
9. It's a beautiful shot.,
10. What a gorgeous field of flowers.,
11. Lovely colors and composition.,
12. I like the contrast of the yellow flowers against the blue sky.,
13. I can almost smell the fresh air.,
14. I love the colors in the sky.,
15. I like the depth of field.,
16. I like the simplicity of this shot.,
17. I love the contrast between the flowers and the sky.,
18. I like the contrast between the flowers and the sky.,
19. I love how the yellow stands out against the blue sky.,
20. Seems a bit oversharpened.,
21. I 'd love to see this in black and white.,
22. I've never seen a field like that before.,
23. I 'd like to see more of the field.,
24. The contrast between the yellow field and the blue sky is great.

Clip selector

1. Lovely shot!,
2. Lovely colours.,
3. Nice composition.,
4. What a beautiful scene!,
5. Beautiful!,
6. Beautiful landscape.,
7. Beautiful sky.,
8. Looks like a painting.,
9. I love the colors and the contrast between the yellow and blue.,
10. This is a lovely shot.,
11. What a gorgeous field of flowers!,
12. Lovely colors and composition.,
13. I love the contrast between the yellow of the field and the blue of the sky.,
14. I can almost smell the fresh air.,
15. I love the colors in this photo.,
16. I love the depth of field.",
17. I like the composition of this shot.,
18. I love the contrast between the field and sky.,
19. I like the contrast between the field and the sky.,
20. I love how the yellow stands out against the blue sky.,
21. Seems a bit oversharpened.,
22. I 'd love to see this in black and white.,
23. I've never seen a field like that before.,
24. I 'd like to see more of the field.,
25. The contrast between the yellow field and the blue sky is great."

Figure 5: Traditional: The sentences are ranked based on the generator's confidence without using the DACS. ARS selector: *ARS* is used to construct the DACS and pick the generated sentences according to *ARS* values. CLIP selector: the fine tuned CLIP model is used to rank the generated sentences to construct the DACS and pick the sentences according to the CLIP matching scores. Note CLIP is only used to pick the best sentence in a group as one of the output sentences of DACS. It is seen that with DACS, the captions are more diverse. With DACS, the captions are more diverse. More example results are available in the Supplementary materials.

and effectiveness of introducing the aesthetically relevant score (*ARS*) for aesthetics image captioning. Even though these metrics do not directly measure aesthetic relevance, the better performances of the new method nevertheless demonstrate the soundness of the new algorithm design.

As described in the main method, with the introduction of *ARS*, we can use *ARS* based diverse aesthetic caption selector (DACS) to generate a diverse set of image captions rather than being restricted to output only one single sentence as in previous methods (Chang, Lu, and Chen 2017). Figures 5 shows examples of aesthetic captions generated with *ARS* based and CLIP based DACS. More examples are available in the Supplementary materials.

Methods	M ↑	R(L) ↑	C↑	S↑
Baseline	0.1227	0.3543	0.0580	0.0172
ARIC (2)	0.1389	0.3610	0.0633	0.0353

Table 6: Image aesthetic captioning performances of our new model as compared with the baseline model.

Image AQA based on Generated Captions

In this experiment, we evaluate image AQA performances based on the generated image captions and results are shown in Table 5. It is seen that using captions selected by the new diverse aesthetic caption selector (DACS) either based on ARS or CLIP performs better than using captions without selection. It is interesting to observe that in multi-modal AQA, including the generated texts has started to slightly exceed image only AQA, but is still quite far from those using ground truth. For example, using ground truth text, the ACC of multi-modal AQA is 0.8407 whilst that using generated captions is 0.8203. Would a future image captioning model be able to generate captions to close the gap? This would be a very interesting question and a goal for future research.

Concluding Remarks

In this paper, we have attempted to study two closely related subjects of aesthetic visual computing, image aesthetic quality assessment (AQA) and image aesthetic captioning (IAC). We first introduce the concept of aesthetic relevance score (ARS) and use it to design the aesthetically relevant image captioning (ARIC) model through an ARS weighted loss function and an ARS based diverse aesthetic caption selector (DACS). We first introduce the concept of aesthetic relevance score (ARS) and have presented extensive experimental results which demonstrate the soundness of the ARS concept and the effectiveness of the ARIC model. We have also contributed a large research database DPC2022 that contains images with both comments and aesthetic ratings.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under grants U22B2035 and 62271323, in part by Guangdong Basic and Applied Basic Research Foundation under grant 2021A1515011584, and in part by the Shenzhen Research and Development Program under grants JCYJ20220531102408020 and JCYJ20200109105008228).

References

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, 382–398. Springer.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6077–6086.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Chang, K.-Y.; Lu, K.-H.; and Chen, C.-S. 2017. Aesthetic critiques generation for photos. In *Proceedings of the IEEE International Conference on Computer Vision*, 3514–3523.

Chen, J.; Guo, H.; Yi, K.; Li, B.; and Elhoseiny, M. 2022. VisualGPT: Data-Efficient Adaptation of Pretrained Language Models for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18030–18040.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Ghosal, K.; Rana, A.; and Smolic, A. 2019. Aesthetic image captioning from weakly-labelled photographs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 0–0.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hossain, M. Z.; Sohel, F.; Shiratuddin, M. F.; and Laga, H. 2019. A Comprehensive Survey of Deep Learning for Image Captioning. *ACM Comput. Surv.*, 51(6).

Jin, X.; Wu, L.; Zhao, G.; Li, X.; Zhang, X.; Ge, S.; Zou, D.; Zhou, B.; and Zhou, X. 2019. Aesthetic attributes assessment of images. In *Proceedings of the 27th ACM International Conference on Multimedia*, 311–319.

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746–1751. Doha, Qatar: Association for Computational Linguistics.

Lai, S.; Xu, L.; Liu, K.; and Zhao, J. 2015. Recurrent Convolutional Neural Networks for Text Classification. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, 2267–2273. AAAI Press. ISBN 0262511290.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Murray, N.; Marchesotti, L.; and Perronnin, F. 2012. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2408–2415. IEEE.

Nieto, D. V.; Celona, L.; and Fernandez-Labrador, C. 2022. Understanding Aesthetics with Language: A Photo Critique Dataset for Aesthetic Assessment. *arXiv e-prints*, arXiv–2206.

Olive, J.; Christianson, C.; and McCary, J. 2011. *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. Springer Publishing Company, Incorporated, 1st edition. ISBN 1441977120.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Pérez, J. M.; Giudici, J. C.; and Luque, F. 2021. pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks. *arXiv preprint arXiv:2106.09462*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28: 91–99.

Valenzise, G.; Kang, C.; and Dufaux, F. 2022. Advances and challenges in computational image aesthetics. *Human Perception of Visual Information*, 133–181.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3156–3164.

Zhang, X.; Gao, X.; He, L.; and Lu, W. 2021. MSCAN: Multimodal Self-and-Collaborative Attention Network for image aesthetic prediction tasks. *Neurocomputing*, 430: 14–23.

Zhang, X.; Gao, X.; Lu, W.; He, L.; and Li, J. 2020. Beyond vision: A multimodal recurrent attention convolutional neural network for unified image aesthetic prediction tasks. *IEEE Transactions on Multimedia*, 23: 611–623.

Zhao, S.; Yao, X.; Yang, J.; Jia, G.; Ding, G.; Chua, T.-S.; Schuller, B. W.; and Keutzer, K. 2021. Affective Image Content Analysis: Two Decades Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1.

Zhou, Y.; Lu, X.; Zhang, J.; and Wang, J. Z. 2016. Joint Image and Text Representation for Aesthetics Analysis. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, 262–266. New York, NY, USA: Association for Computing Machinery. ISBN 9781450336031.