

# To be an Artist: Automatic Generation on Food Image Aesthetic Captioning

Xiaohan Zou, Cheng Lin, Yinjia Zhang, Qinpei Zhao\*

*School of Software Engineering*

*Tongji University*

*Shanghai, China*

qinpeizhao@tongji.edu.cn

**Abstract**—Image aesthetic captioning is a multi-modal task that is to generate aesthetic critiques for images. In contrast to common image captioning tasks, where different captions aimed at providing factual descriptions of a same image are always similar, captions with respect to different aesthetic attributes of the same image can be totally different in an aesthetic captioning task. Such inter-aspect differences are always overlooked, which leads to the lack of diversity and coherence of the captions generated by most of the existing image aesthetic captioning systems. In this paper, we propose a novel model to generate aesthetic captions for food images. Our model redefines food image aesthetic captioning as a compositional task that consists of two separated modules, i.e., a single-aspect captioning and an unsupervised text compression. The first module is guaranteed to generate the captions and learn feature representations of each aesthetic attribute. Then, the second module is supposed to study the associations among all feature representations and automatically aggregate captions of all aesthetic attributes to a final sentence. We also collect a dataset which contains pair-wise image-comment data related to six aesthetic attributes. Two new evaluation criteria are introduced to comprehensively assess the quality of the generated captions. Experiments on the dataset demonstrate the effectiveness of the proposed model.

**Index Terms**—image aesthetic captioning, aesthetic critiques, text generation, image-comment data

## I. INTRODUCTION

Image aesthetic assessment has been an important topic in the field of computer vision in the last decades [1–3]. Research on this topic is supposed to help people with creating, selecting and sharing digital images. When the task is specified to assessing visual aesthetics of food images, its goals also include helping businesses such as *Yelp* to create food images with higher aesthetic quality for advertising so that they can attract more customers.

In most of the literature, the problem of computing image aesthetic is formulated as a classification or regression problem [1–3], the purpose of which is to classify images into a high aesthetic quality category and low aesthetic quality category or give images numerical scores based on their aesthetic quality.

However, an image is worth thousands of words. Numerical scores are not able to describe them comprehensively. When assessing image quality, instead of giving a score, humans are always more willing to provide some captions to review given images from several aesthetic aspects like color, lighting, composition, etc. Chang et al. [4] were the first to address the image aesthetic captioning problem and opened up a new

area of research of image aesthetic computing. They also constructed Photo Critique Captioning Dataset (*PCCD*) for this task, which contains about 4,200 images from professional photo critique website *GuruShots*, and each image is attached with comments of 7 aesthetic attributes. Unlike datasets for common image captioning tasks in which different captions aimed at providing factual descriptions of the same image are always similar, captions with respect to different aesthetic attributes of a same image in datasets for image aesthetic captioning like *PCCD* can be totally different. Hence, simply applying methods for image captioning task to generate image aesthetic captions is likely to lead to the lack-of-diversity problem of the generated captions because of the inter-attribute differences between the words and sentence forms of the captions, which is also observed by the work in [4]. However, although they considered merging captions from several aesthetic aspects and generating abundant aesthetic captions via an additional language model with an attention network for solving this problem, the outputs of their model are still lack of diversity, most of which only describe one aesthetic attribute. This is because they still use single-aspect captions as the targets of the new language model, due to the lack of ground truth multi-attribute captions.

Another idea is to produce captions for each aesthetic attribute of an image to ensure a full review of aesthetic attributes in [5]. However, there is little continuity between the captions of each aesthetic aspect and most people are more willing to read a brief and coherent caption. Also, the associations between sentences of different aspects are seldom learned.

To address the above-mentioned issues, in this work, we proposed a novel method which consists of two modules, i.e., single-aspect captioning and unsupervised text summarization, as illustrated in Fig. 1. The first module is guaranteed to generate captions for each aesthetic aspect, which is based on the common framework that combines Convolutional Neural Networks (*CNNs*) with Recurrent Neural Networks (*RNNs*) for image captioning. The second module can be regarded as an unsupervised text summarization model based on a Denoising Auto-Encoder (*DAE*). It is guaranteed to summarize captions from all aesthetic aspects to a final sentence. In addition, we build a new dataset that contains food images and their corresponding comments, where an image could have

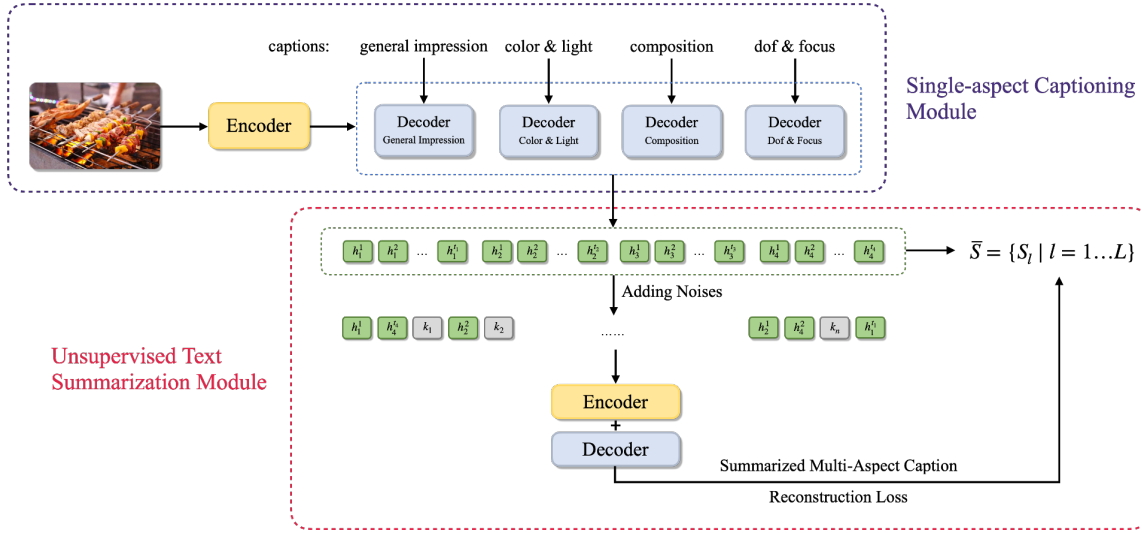


Fig. 1. An overview of the proposed model. The single-aspect captioning module generates captions and extracts feature representations of each attribute aspect. Then the unsupervised text summarization module fuses them to a comprehensive and integrated sentence as the final aesthetic caption.

multiple comments related to up to 6 aspects of the aesthetics. There are 7,172 images and 46,485 comments in total. We then comprehensively evaluate our method of single-aspect captioning and multi-attribute captioning on the introduced dataset using both the existing image captioning metrics and two introduced evaluation metrics.

The contributions of this paper are summarized as follows:

- We propose a novel compositional framework consisting of a single-aspect captioning module and an unsupervised text summarization module for generating comprehensive and coherent aesthetic captions for food images.
- We introduce a dataset with 7,172 food images and 46,485 comments for food images aesthetic captioning, which contains captions of up to 6 aesthetic attributes of the images.

## II. RELATED WORK

### A. Image Captioning

Most of the image captioning models adopt a *CNN-RNN* framework and achieve promising results [6, 7]. Most of them introduce the attention mechanism. Xu et al. [8] introduce a soft attention mechanism, which enables the model to focus on different regions of the image when generating different words. Lu et al. [9] extend the attention module with the ability of deciding when to look at the image and when to rely on the language model while generating words. Anderson et al. [10] combine bottom-up and top-down attention mechanisms and enable attentions to be calculated at salient image regions.

There is also a trend of producing non-factual descriptions for images. Mathews et al. [11] and Zhao et al. [12] generate captions with sentiments. Gan et al. [13] produce humorous and romantic captions. Mathews et al. [14] generate a story about the associated image. However, all these work focus on stylised descriptions instead of descriptions related to art and aesthetics.

### B. Image Aesthetic Captioning

Research in the area of generating captions related to image aesthetics is still rare. This problem was first addressed by Chang et al. [4]. They apply a traditional CNN-LSTM model to extract features from the data of three different aesthetic aspects and then generate meaningful captions by fusing them together. However, the captions generated by their model are still not comprehensive, because of the lack of multi-attribute reference captions in the training data. Jin et al. [5] produce scores and captions for each aspect using a multi-task network to ensure that a full review of the aesthetic attributes is given. However, there is little continuity and coherence between the captions of different aspects output by their method. Wang et al. [15] produce aesthetic scores and captions regardless of the aesthetic aspect. Ghosal et al. [16] propose a strategy for filtering aesthetic captions when building a dataset and a weakly-supervised approach for training the CNN. They ignore the lack-of-diversity problem of the generated captions. Also, as food images are a sub-category of the images, there is rare work on the food image aesthetic captioning.

## III. PROPOSED MODEL

Assume that the training data has been separated into different aspects. Namely, a dataset  $D = (I_i, S_i, a_i), i = 1 \dots N$ , is available to train our food image aesthetic captioning model, where  $I_i$  is the  $i$ -th image,  $S_i$  is the caption of  $I_i$ ,  $a_i \in \{1 \dots L\}$  is the aesthetic aspect attached to caption  $S_i$  and  $L$  is the number of the aspects.

As shown in Fig 1, the proposed framework is divided into two parts: the single-aspect captioning module and the unsupervised text summarization module. The first module generates captions and extracts feature representations of each aspect. It is achieved by a CNN based encoder followed by a two-layer LSTM based decoder with an adaptive attention mechanism and a look-back mechanism. The second module

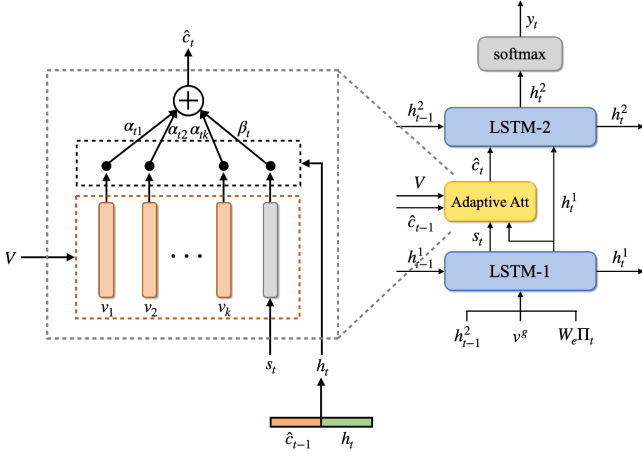


Fig. 2. An illustration of the decoder of our single-aspect captioning model generating the word  $y_t$

aggregates the captions from all aspects and forms a sentence as the final aesthetic caption. It is a denoising auto-encoder consisting of a bidirectional LSTM as the encoder and a LSTM as the decoder.

#### A. Single-aspect Captioning

In our single-aspect captioning module, we train the image captioning model for each single aspect. Namely,  $(I_i, S_i, l), i = 1 \dots N_l$  are used for each model, where  $N_l$  is the amount of training data of aspect  $l$ . An encoder-decoder framework is employed for generating captions.

Given an image  $I_i$ , a CNN based encoder is used to extract  $k$  feature vectors  $V = \{v_1, v_2, \dots, v_k\}, v_i \in \mathbb{R}^D$ , each of which is a  $D$ -dimensional representation corresponding to a part of the image. In our implementation,  $V$  is the  $7 \times 7 \times 2048$  ( $k = 49, D = 2048$ ) feature map of the last convolutional layer of a ResNet-101 [17] CNN pretrained on ImageNet [18]. The mean pooling vector  $v^g$  is taken as the global image feature.

Inspired by [10], our captioning model is composed of two standard LSTM layers, each of them computes the hidden state  $h_t$  at time  $t$  using a recurrence formula:

$$h_t = \text{LSTM}(x_t, h_{t-1}) \quad (1)$$

where  $x_t$  is the input vector. The decoder of the captioning model is illustrated in Fig 2.

**Attention LSTM.** The first LSTM (*LSTM-1*) is taken as a visual attention model, which is to learn the contextual information and guide the attention module to select the highlighting image regions. The second LSTM (*LSTM-2*) is characterized as a language model.

The input of the attention LSTM at each time step consists of the previous hidden state of the language LSTM, the global image feature  $v^g$  and the current input word:

$$x_t^1 = [h_{t-1}^2; v^g; W_e \Pi_t] \quad (2)$$

where  $W_e$  is the word embedding matrix and  $\Pi_t$  is one-hot encoding of the input word at time  $t$ . The input provides the attention LSTM with maximum contextual information.

**Look Back.** Conventional attention module calculates the context vector  $\hat{c}_t$  by:

$$\hat{c}_t = f_{att}(V, h_t) \quad (3)$$

where  $f_{att}$  is the attention network. However, the attention region should have visual coherence. The visual information of the the previous time step  $\hat{c}_{t-1}$  can also contribute to the current attention generation [19]. So we concatenate  $\hat{c}_{t-1}$  with the current hidden state  $h_t$  as the input of  $f_{att}$ :

$$\hat{h}_t = [h_t; \hat{c}_{t-1}] \quad (4)$$

**Adaptive Attention.** Given the spatial image feature  $V \in \mathbb{R}^{k \times D}$  and the concatenated input  $\hat{h}_t$ , the attention weights  $\alpha \in \mathbb{R}^k$  over the  $k$  regions are calculated as follows:

$$z_t = w_h \tanh(W_v V + W_g \hat{h}_t) \quad (5)$$

$$\alpha_t = \text{softmax}(z_t) \quad (6)$$

where  $W_v, W_g$  and  $w_h$  are learnable weight parameters. Then the visual context vector  $c_t$  is generated by:

$$c_t = \sum_{i=1}^k \alpha_{ti} v_i \quad (7)$$

However, not all words in the caption have corresponding visual information. Some non-visual words can be predicted reliably just using linguistic information. Therefore, here we adopt the adaptive attention module proposed in [9]. We first backup the latent representation of the attention LSTM's memory by:

$$g = \sigma(W_x x_t^1 + W_h h_{t-1}^1) \quad (8)$$

$$s_t = g_t \odot \tanh(m_t^1) \quad (9)$$

where  $W_x$  and  $W_h$  are learnable weight parameters,  $\sigma$  is the sigmoid activation function,  $m_t^1$  is the cell state of the attention LSTM and  $\odot$  represents the hadamard product.  $g_t$  is a gate to determine how much memory should be backup.

Then a scalar  $\beta$  is computed to determine how much visual information and how much linguistic information should we take when generating the next word:

$$\hat{\alpha}_t = \text{softmax}([z_t; w_h \tanh(W_s s_t + W_g \hat{h}_t)]) \quad (10)$$

$$\beta = \hat{\alpha}_t[k+1] \quad (11)$$

where  $W_s, W_g$  and  $w_h$  are learnable weight parameters. Notably,  $W_g$  and  $w_h$  are the same weight parameters as in Eq. 5.  $\beta$  is in the range  $[0, 1]$ ,  $\beta = 1$  implies that only the linguistic information is used and  $\beta = 0$  means only the visual information is used.

The final context vector  $\hat{c}_t$  can be obtained by:

$$\hat{c}_t = \beta_t s_t + (1 - \beta_t) c_t \quad (12)$$

**Language LSTM.** The language LSTM is supposed to generate the final captions for each aesthetic aspect. Its input consists of the context vector  $\hat{c}_t$  and the hidden state of the attention LSTM, given by:

$$x_t^2 = [\hat{c}_t; h_t^1] \quad (13)$$

So the probability over a vocabulary of possible words at time  $t$  can be calculated as:

$$p(y_t | y_{1:t-1}) = \text{softmax}(W_p h_t^2 + b_p) \quad (14)$$

where  $W_p$  and  $b_p$  are learnable weights and biases, and  $y_{1:T}$  refers to a sequence of words  $(y_1, y_2, \dots, y_T)$ .

In the training phase with cross-entropy loss, given the model parameters  $\theta$  and the correct caption  $y_{1:T}^*$ , the loss function can be defined as:

$$L(\theta) = - \sum_{t=1}^T \log(p_\theta(y_t^* | y_{1:t-1}^*)) \quad (15)$$

### B. Unsupervised Text Summarization

Our unsupervised text summarization module is guaranteed to summarize the captions generated by  $L$  single-aspect captioning models to a coherent multi-attribute sentence. Due to the lack of ground truth multi-attribute captions, we treat this task as an unsupervised learning problem and adopt a denoising auto-encoder model to achieve it [20].

A straightforward way is to embed the word sequence generated by each single-aspect captioning model and feed the word embedding vectors into the summarization model. However, it is found that directly feeding the hidden states output by  $L$  single-aspect captioning models to the summarization model can achieve better results. This is probably because both images and sentences are used as the inputs during the learning process of single-aspect captioning models, so the hidden states can serve as the deep feature representations extracted by these models for each aspect. Thus, they are better sources for training the summarization model and make it more effective in generating diverse and suitable captions.

We denote the hidden states and word sequence output by the language LSTM layer of the  $l$ -th single-aspect captioning model to be  $\mathbf{h}_l = \{h_{lt}^2 | t = 1 \dots T_l\}$  and  $S_l = \{S_{lt} | t = 1 \dots T_l\}$ . So the hidden states and word sequences of all  $L$  aspect models can be denoted as  $H = \{\mathbf{h}_l | l = 1 \dots L\}$  and  $\bar{S} = \{S_l | l = 1 \dots L\}$ .

The core of our denoising auto-encoder model is based on the framework introduced in [21]. It consists of a bidirectional LSTM as the encoder and an attentional LSTM as the decoder.

**Adding Noises.** Firstly, we add noises to the original sequence  $H$ . Then, we train our text summarization model to reconstruct the original sentence  $\bar{S}$ . Such a strategy forces the model to remove and reorder the elements and thus learn

how to generate short but correct and coherent captions. In this way, we can get rid of the dependence on paired data.

There are many ways of adding noises, like deleting and shuffling the words [22] or replacing some of the words by other strings [23]. Inspired by these works, we randomly extend and shuffle the original sequence.

We randomly sample additional words from the vocabulary and append them to the original sequence  $H$ . As shown in Fig 1,  $k_i = W_k \Pi_i$  is the embedding vector of the  $i$ -th sampled word, where  $W_k$  is the embedding weight and  $\Pi_i$  is one-hot encoding of the word. In practice, our goal is to extend 50% of the original sequence. Next, we shuffle the extended sequence. The noised sequence is noted as  $\tilde{H}$ .

**Encoder.** The encoder  $\phi^{\text{TS-E}}$  of our summarization model is based on a bidirectional LSTM. It reads the input sequence  $\tilde{H}$  from both directions and computes hidden states for each element:

$$\vec{h}_i = \overrightarrow{\text{LSTM}}(\vec{h}_{i-1}, \tilde{H}_i) \quad (16)$$

$$\overleftarrow{h}_i = \overleftarrow{\text{LSTM}}(\overleftarrow{h}_{i-1}, \tilde{H}_i) \quad (17)$$

The final hidden representation of the  $i$ -th element is  $h_i^{\text{TS-E}} = [\vec{h}_i; \overleftarrow{h}_i]$ .

**Decoder.** The decoder  $\phi^{\text{TS-D}}$  is a LSTM attached with an attention network implemented in [21]. The desired output of our summarization model should be shorter than that of the original sequence. Hence, an additional scalar is concatenated with the original input  $x_i^{\text{TS-D}}$  to induce the model to generate the sentence of a desired length:

$$\hat{x}_i^{\text{TS-D}} = [x_i^{\text{TS-D}}; L_{\text{target}} - i] \quad (18)$$

where  $L_{\text{target}}$  is the desired length and  $i$  is the length of the sequence that has been generated already.  $L_{\text{target}} - i$  falls to 0 when the generated sequence reaches the desired length  $L_{\text{target}}$ , and then goes negative.

This module is trained with a cross-entropy loss to reconstruct the original sentences by minimizing the reconstruction loss. Given the model parameters  $\theta^{\text{TS}}$ , the reconstruction loss can be defined as:

$$L(\theta^{\text{TS}}) = - \sum_{i=1}^{T_{\bar{S}}} L_{\text{cross-entropy}}(\bar{S}_i, \phi^{\text{TS-D}}(\phi^{\text{TS-E}}(\bar{S}_i))) \quad (19)$$

In this module, we adopt an unsupervised method to fuse the captions from all different aspects, which gets rid of the paired data and the lack-of-diversity problem suffered by [4].

## IV. DATASET AND EVALUATION CRITERIA

### A. Dataset

All of the current existed datasets for the aesthetic captioning task (PCCD [4], AVA-Reviews [15], AVA-Captions [16] and DPC-Captions [5]) are for general pictures, regardless of their subjects. However, from a photographic point of view, the standards for assessing food images and other subjects are different. Therefore, we construct a dataset to validate the

proposed method on the task specified to food images. The dataset is based on the images together with their comments crawled from the website of DPChallenge.com<sup>1</sup>. The raw data contains 7,312 images and more than 83,000 comments.

For the comments, the preprocessing is essential. We firstly handle the generic noises, such as punctuation, emoji, non-English comments, abbreviation (such as “imo” = “in my opinion”) and exclamatory words (such as “niceeeeeee”), using a standard natural language processing toolkit [24].

Then the comments providing valid but less informative description like “*Love the composition.*”, which we refer to as *safe* comments, are filtered following the work [16]. Otherwise, the model may learn these less-informative captions instead of the more informative ones like “*This is well done, like the grain, shadows, the placement of the pears and how the color of the wall matches the colors on the pears.*”.

The informativeness score of a caption is computed by:

$$\text{Score} = -\frac{1}{2} \left[ \log \prod_i^N P(u_i) + \log \prod_j^M P(b_j) \right] \quad (20)$$

where  $S_u = (u_1, u_2, \dots, u_N)$  is the set of unigrams (only nouns are selected) and  $S_b = (b_1, b_2, \dots, b_M)$  is the set of bigrams (only “descriptor-object” patterns are selected).  $P(\cdot)$  computes the corpus probability for the given n-gram:

$$P(\omega) = \frac{C_\omega}{\sum_{i=1}^D C_i} \quad (21)$$

where  $D$  is the vocabulary size and  $C_\omega$  is the frequency of n-gram  $\omega$ . The scoring strategy is motivated by the commonly used TF-IDF method, which believes the key information in a sentence is stored in the low-frequency n-grams. So a sentence composed of frequently occurring n-grams such as “color” or “composition” is less likely to contain useful information than the one composed of seldom occurring n-grams like “simplicity” or “bottom half”. Examples on comments with informativeness scores are shown in Fig. 3.

Those comments with an informativeness score lower than the threshold are removed, which is set as 15 in our implementation. We then remove the images which are left with no informative comments.

Then these comments are classified into 6 aesthetic aspect (*Color & Light*, *Composition*, *Dof & Focus*, *Subject*, *Use of Camera* and *General Impression*) using the LDA (Latent Dirichlet Allocation) model. LDA is an unsupervised method and widely used for topic modelling. In our case, it generates a topic attached to some keywords for each comment. Then, each comment is assigned an aesthetic aspect manually based on its topic keywords. We remove the comments which can not be classified into any of the aspect.

Finally, 7,172 images and 46,485 comments are left, which means 44% of the comments are filtered by our cleaning strategy. Table I shows the statistics of our dataset.

<sup>1</sup><https://www.dpchallenge.com/>


Images	Comments	Scores
	Good picture.	5.98
	Very funny, and great color.	9.50
	Cute idea, but also well executed, the composition and lighting are both great.	17.45
	Your cropping is a little tight on the right and there may be some motion blur. The shot isn't bad, it just may need a different crop or some fill light on the right.	65.60

Fig. 3. Sample examples on the comments with informativeness scores.

TABLE I  
STATISTICS OF OUR FOOD IMAGE AESTHETIC CAPTIONING DATASET

Aesthetic Aspect	# Images	# Captions	Vocabulary Size
Color & Light	4,708	9,402	9,425
Composition	2,778	4,092	5,802
Dof & Focus	3,843	6,758	7,069
Subject	4,104	7,706	9,031
Use of Camera	1,188	1,487	3,766
General Impression	5,835	17,040	12,859
Total	7,172	46,485	20,987

## B. Evaluation Criteria

The automatic evaluation of the food image aesthetic captioning task becomes also an issue since the output is highly flexible [4]. The criteria to evaluate our models include *BLEU* [25], *METEOR* [26], *CIDEr* [27], *ROUGE-L* [28], and *SPICE* [29], which are commonly used in nature language generation problems. The BLEU, METEOR, CIDEr and ROUGE-L compute n-gram overlaps to perform the evaluation. In contrast to them, the SPICE parses the generated and reference sentences into scene graphs first and then evaluates the similarity between the parsed graphs. It reports the *F-score* as the evaluation results. It has been shown that the SPICE captures human judgments and evaluates semantic similarity better than other metrics.

However, it is unreasonable to only use these traditional metrics for evaluating this extremely flexible task. There are multiple ideal aesthetic captions for a given picture. Hence we also develop two evaluation metrics to comprehensively measure the quality of the captions.

**Diversity.** Image aesthetic captioning models often suffer from the problem caused by valid but monotonous captions. For example, “*I love the color.*” repeated for every picture. These captions have little reference value and make people feel tedious. However, such problem cannot be reflected by the evaluation metrics mentioned above. Therefore, we calculate the similarity between the generated sentence  $a$  and  $b$  by the *Jaccard* similarity function:



$$sim_n(a, b) = \frac{|g_n(a) \cap g_n(b)|}{|g_n(a)| + |g_n(b)| - |g_n(a) \cap g_n(b)|} \quad (22)$$

where  $g_n(\cdot)$  is the function that returns the set of all n-grams in the given sentence. We treat sentence  $a$  and  $b$  different if  $sim_n(a, b)$  is smaller than a threshold (in our implementation 30% is used). Finally, we report the proportion of non-duplication sentences as the diversity of the outputs.

**Novelty.** We hope the model is able to generate novel and integrated captions. Hence, we also calculate the novelty of the outputs by computing the difference between the generated sentence  $c_i$  and the corresponding sentences  $S_i$  in the training data:

$$N(c_i, S_i) = \frac{|g_n(c_i) \cap g_n(S_i)|}{|g_n(c_i)| + |g_n(S_i)| - |g_n(c_i) \cap g_n(S_i)|} \quad (23)$$

## V. EXPERIMENTS

A series of experiments are conducted to evaluate the effectiveness of the proposed models based on the newly collected dataset.

### A. Single-aspect Captioning

**Baseline.** We compare our single-aspect captioning model (referred as SAC) with three baseline approaches:

- **img2txt.** The Google’s img2txt model [6] is based on a basic encoder-decoder architecture. It is worth noting that the original model uses Inception V3 [30] for extracting image features. However, we change it to ResNet-101 in the experiment for fairness.
- **Soft Attention.** This model is based on the Soft Attention model [8]. The original model uses VGGNet [31] for encoding, so we also change it to ResNet-101.
- **Adaptive Attention.** This model is based on the Adaptive Attention model [9].

The comparison results are shown in Table II. The results show that our approach achieves the best performance in all metrics and for each aesthetic aspect, which demonstrates the effectiveness of our approach. Some examples of the captions from our module are shown in Fig. 4.

**Ablation Study.** To understand the importance of the key components in the proposed model, an ablation study is performed by training multiple ablated versions of the model:

- **SAC: Soft Attention.** Replacing the adaptive attention module by a soft attention module.
- **SAC: Single LSTM.** Using a single LSTM layer instead of two LSTM layers in the decoder.
- **SAC: No Look Back.** Without look back mechanism.
- **SAC: Full Model.** The full model proposed in Section III-A.

The evaluation results are shown in Table III. The results show that all the three methods bring with noteworthy improvement in various aspects, which indicates that all components of our approach contribute to improve the quality of the generated captions.





Images	Captions
	<b>Color &amp; Light:</b> I think this would have been more effective if the lighting was a bit more.
	<b>Composition:</b> I really like the idea, but I think it would have been better if the glass was in the background.
	<b>General Impression:</b> I really like the idea of this shot.
	<b>Dof &amp; Focus:</b> I think the dof is a little bit shallow and the background is a little bit distracting.

Fig. 4. Examples of the captions generated by our single-aspect captioning module for different aspects.

We also find that the contribution of the improvement on the language model is higher than the improvement on attention mechanism in this task. It is probably because the generated captions in this task are more related to the aesthetic features than the semantic information of the images, so the performance relies less on the attention module.

### B. Multi-aspect Captioning

For multi-aspect captioning, the aggregated sentence could be quite different from different methods. We compare our model (referred as *SACTC*) with two baseline approaches:

- **img2txt.** We directly apply the model proposed in [6] to the whole dataset which contains the training captions from all aspects. Its original encoder is replaced by a ResNet-101 model.
- **SAC.** We directly apply our single-aspect captioning model (introduced in Section III-A) to the whole dataset.

**Ablation Study.** We also perform an ablation study:

- **SACTC: Word Embeddings.** Take the word embedding vectors of the sequences output by single-aspect captioning models as the input of the summarization model.
- **SACTC: Hidden States.** The proposed method, directly feeds the hidden states output by single-aspect captioning models into the summarization model.

**Results.** Due to the lack of multi-aspect captions as reference in our dataset, we are not able to evaluate the model using traditional metrics. Hence here we use the diversity and

TABLE II  
THE COMPARISONS ON THE PERFORMANCE OF DIFFERENT MODELS ON EACH AESTHETIC ASPECT.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr	METEOR	SPICE
img2txt (General Impression)	47.3	27.1	14.0	6.8	23.8	6.0	11.7	17.9
Soft Attention (General Impression)	47.7	27.5	14.7	7.1	26.4	6.4	12.1	18.4
Adaptive Attention (General Impression)	47.9	27.8	14.8	7.1	24.3	6.5	12.4	18.7
<b>SAC (General Impression)</b>	<b>49.4</b>	<b>28.3</b>	<b>15.9</b>	<b>7.6</b>	<b>25.4</b>	<b>7.0</b>	<b>13.5</b>	<b>19.6</b>
img2txt (Color & Light)	44.3	24.2	14.5	6.2	24.8	5.9	10.6	14.6
Soft Attention (Color & Light)	45.0	24.6	14.7	6.8	25.3	6.1	11.1	15.5
Adaptive Attention (Color & Light)	46.1	25.1	15.0	6.9	25.6	6.0	11.4	15.8
<b>SAC (Color &amp; Light)</b>	<b>49.7</b>	<b>27.9</b>	<b>15.6</b>	<b>7.2</b>	<b>26.4</b>	<b>6.4</b>	<b>12.7</b>	<b>17.7</b>
img2txt (Composition)	45.2	23.5	13.4	6.3	24.3	6.0	11.5	17.2
Soft Attention (Composition)	46.0	23.9	13.8	6.4	24.9	6.3	11.8	17.6
Adaptive Attention (Composition)	46.4	24.1	14.0	6.6	24.8	6.4	12.0	18.0
<b>SAC (Composition)</b>	<b>48.6</b>	<b>25.6</b>	<b>14.9</b>	<b>7.0</b>	<b>25.9</b>	<b>6.8</b>	<b>13.0</b>	<b>18.2</b>
img2txt (Dof & Focus)	44.8	23.4	13.2	6.0	24.8	5.2	10.3	14.9
Soft Attention (Dof & Focus)	45.7	24.0	13.7	6.5	25.3	5.8	10.8	15.4
Adaptive Attention (Dof & Focus)	45.4	23.8	13.5	6.3	25.6	5.6	11.0	15.3
<b>SAC (Dof &amp; Focus)</b>	<b>46.8</b>	<b>24.9</b>	<b>14.3</b>	<b>6.7</b>	<b>26.4</b>	<b>6.2</b>	<b>12.3</b>	<b>17.0</b>

The BLEU-1,2,3,4, ROUGE-L, CIDEr, METEOR and SPICE are reported. All values refer to percentage (%). The proposed model and the best performance is highlighted in bold.

TABLE III  
ABLATION STUDY ON EACH AESTHETIC ASPECT

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr	METEOR	SPICE
SAC: Soft Attention (General Impression)	48.7	27.9	15.3	7.4	25.0	6.7	12.7	18.9
SAC: Single LSTM (General Impression)	48.2	27.6	15.1	7.2	24.7	6.7	12.4	19.4
SAC: No Look Back (General Impression)	49.0	28.0	15.6	7.5	25.2	6.9	13.2	19.2
<b>SAC: Full Model (General Impression)</b>	<b>49.4</b>	<b>28.3</b>	<b>15.9</b>	<b>7.6</b>	<b>25.4</b>	<b>7.0</b>	<b>13.5</b>	<b>19.6</b>
SAC: Soft Attention (Color & Light)	47.3	26.8	15.1	7.1	25.8	6.2	12.1	16.2
SAC: Single LSTM (Color & Light)	45.5	25.6	14.8	6.9	25.5	6.2	11.8	17.4
SAC: No Look Back (Color & Light)	48.8	27.4	15.3	7.2	26.2	<b>6.5</b>	12.4	16.8
<b>SAC: Full Model (Color &amp; Light)</b>	<b>49.7</b>	<b>27.9</b>	<b>15.6</b>	<b>7.2</b>	<b>26.4</b>	6.4	<b>12.7</b>	<b>17.7</b>
SAC: Soft Attention (Composition)	47.6	24.7	14.5	6.8	25.2	6.7	11.9	17.9
SAC: Single LSTM (Composition)	46.9	24.3	14.2	6.6	25.0	6.4	12.1	<b>18.7</b>
SAC: No Look Back (Composition)	48.1	25.1	14.6	6.9	25.5	6.2	12.6	18.4
<b>SAC: Full Model (Composition)</b>	<b>48.6</b>	<b>25.6</b>	<b>14.9</b>	<b>7.0</b>	<b>25.9</b>	<b>6.8</b>	<b>13.0</b>	18.2
SAC: Soft Attention (Dof & Focus)	46.1	24.5	14.0	6.6	25.8	5.7	11.4	16.0
SAC: Single LSTM (Dof & Focus)	45.9	24.2	13.8	6.5	25.5	5.8	11.7	16.5
SAC: No Look Back (Dof & Focus)	46.3	24.6	14.1	6.7	26.2	6.0	12.1	<b>17.2</b>
<b>SAC: Full Model (Dof &amp; Focus)</b>	<b>46.8</b>	<b>24.9</b>	<b>14.3</b>	<b>6.7</b>	<b>26.4</b>	<b>6.2</b>	<b>12.3</b>	17.0

The BLEU-1,2,3,4, ROUGE-L, CIDEr, METEOR and SPICE are reported. All values refer to percentage (%). The full model and the best performance is highlighted in bold.


Image	Aesthetic Captions
	<b>General Impression:</b> the idea is cool i like the great detail and color
	<b>Color &amp; Light:</b> i like the lighting on the cup but I think the background is too bright
	<b>Composition:</b> nice idea but i would like to see more of the cup in the center
	<b>Dof &amp; Focus:</b> this is a great shot but I think the focus is a little bit soft and the background is distracting
	<b>img2txt:</b> nice idea but i think the light from right is a little bit harsh
	<b>SACTC:</b> cool idea and great detail and lighting on the cup but the background is bright and distracting like more cup in center the focus is soft

Fig. 5. An example of multi-aspect captions generated by the img2txt and the proposed model. We also report the captions generated by our single-aspect captioning models in four aesthetic aspects for reference.

TABLE IV  
PERFORMANCE ON MULTI-ASPECT CAPTIONING

Method	D-1	D-4	N-1	N-4	S	B-4
img2txt	69.9	81.7	53.4	62.6	—	—
SAC	71.8	84.5	58.7	67.7	—	—
SACTC: Word Embeddings	93.6	98.2	76.13	<b>86.8</b>	<b>9.8</b>	4.9
<b>SACTC: Hidden States</b>	<b>94.2</b>	<b>98.7</b>	<b>77.71</b>	84.9	9.4	<b>5.1</b>

D- $n$  evaluates the diversity and N- $n$  evaluates the novelty of the generated captions in  $n$ -grams. All values refer to percentage (%). The proposed model and the best performance is highlighted in bold.

novelty metrics introduced in IV-B. We also report the BLEU-4 ( $B-4$ ) and SPICE ( $S$ ) scores computed using the single-aspect captions for reference. Table IV shows the evaluation results.

Results show that our approach outperforms all baselines substantially in terms of diversity and novelty, which indicates our approach’s ability of outputting comprehensive and integrated captions. It is also found that the approach that leverages hidden states of different aspects is superior to the one that uses word embedding vectors as the input. An example of the captions generated by our method and the img2txt is shown in Fig. 5. It is obvious that our method can outperform the baseline in terms of diversity and continuity. The sentence from the img2txt includes only one aesthetic aspect. However, the one from our method includes all of the four aspects. Also, our method is able to capture the semantic associations between captions of different aspects and exclude unimportant phrases such as “*nice idea*”.

## VI. CONCLUSION

In this work, we proposed a novel method consisting of a single-aspect captioning module and an unsupervised text summarization module to generate comprehensive, integrated and coherent aesthetic captions for the food images. A new dataset is built for evaluating the proposed model on this task. Two new evaluation metrics are also introduced to comprehensively assess the quality of the generated aesthetic captions. The experimental results demonstrate the effectiveness of our approach, especially in terms of diversity and novelty.

## REFERENCES

- [1] Datta, Ritendra, et al. Studying aesthetics in photographic images using a computational approach. European conference on computer vision. Springer, Berlin, Heidelberg, 2006.
- [2] Lu, Xin, et al. Rapid: Rating pictorial aesthetics using deep learning. Proceedings of the 22nd ACM international conference on Multimedia. 2014.
- [3] Kao, Yueying, Ran He, and Kaiqi Huang. Deep aesthetic quality assessment with semantic information. IEEE Transactions on Image Processing 26.3 (2017): 1482-1495.
- [4] Chang, Kuang-Yu, Kung-Hung Lu, and Chu-Song Chen. Aesthetic critiques generation for photos. Proceedings of the IEEE International Conference on Computer Vision. 2017.
- [5] Jin, Xin, et al. Aesthetic Attributes Assessment of Images. Proceedings of the 27th ACM International Conference on Multimedia. 2019.
- [6] Oriol Vinyals, Alexander Toshev, Samy Bengio and Dumitru Erhan. Show and tell: A neural image caption generator. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.
- [7] Andrej Karpathy, and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.
- [8] Xu, Kelvin, et al. Show, attend and tell: Neural image caption generation with visual attention. Proceedings of the International Conference on Machine Learning. 2015.
- [9] Jiasen Lu, Caiming Xiong, Devi Parikh and Richard Socher. Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
- [10] Anderson, Peter, et al. Bottom-up and top-down attention for image captioning and visual question answering. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
- [11] Alexander Patrick Mathews, Lexing Xie, and Xuming He. Senticap: Generating image descriptions with sentiments. Thirtieth AAAI Conference on Artificial Intelligence. 2016.
- [12] Wentian Zhao, Xinxiao Wu, and Xiaoxun Zhang. MemCap: Memorizing Style Knowledge for Image Captioning. Thirty-Fourth AAAI Conference on Artificial Intelligence. 2020.
- [13] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao and Li Deng. StyleNet: Generating Attractive Visual Captions with Styles. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
- [14] Alexander Mathews, Lexing Xie, and Xuming He. Semstyle: Learning to generate stylised image captions using unaligned text. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
- [15] Wenshan Wang, Su Yang, Weishan Zhang and Jiulong Zhang. Neural aesthetic image reviewer. IET Computer Vision. 2019.
- [16] Ghosal, Koustav, Aakanksha Rana, and Aljosa Smolic. Aesthetic Image Captioning From Weakly-Labelled Photographs. Proceedings of the IEEE International Conference on Computer Vision Workshops. 2019.
- [17] He, Kaiming, et al. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
- [18] Russakovsky, Olga, et al. Imagenet large scale visual recognition challenge. International journal of computer vision 115.3 (2015): 211-252.
- [19] Qin, Yu, et al. Look back and predict forward in image captioning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.
- [20] Fevry, Thibault, and Jason Phang. Unsupervised sentence compression using denoising auto-encoders. arXiv preprint arXiv:1809.02669 (2018).
- [21] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. International Conference on Learning Representations. 2015.
- [22] Artetxe, Mikel, et al. Unsupervised neural machine translation. International Conference on Learning Representations. 2018.
- [23] Devlin, Jacob, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT). 2019.
- [24] Loper, Edward, and Steven Bird. NLTK: the natural language toolkit. arXiv preprint cs/0205028 (2002).
- [25] Papineni, Kishore, et al. “BLEU: a method for automatic evaluation of machine translation.” Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.
- [26] Satyanjeev Banerjee, and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. Meeting of the Association for Computational Linguistics. 2005.
- [27] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDER: Consensus-based image description evaluation. Conference on Computer Vision and Pattern Recognition. 2015.
- [28] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. Proceedings of Workshop on Text Summarization Branches Out Post Conference Workshop of ACL. 2004.
- [29] Peter Anderson, Basura Fernando, Mark Johnson and Stephen Gould. SPICE: Semantic Propositional Image Caption Evaluation. European Conference on Computer Vision. 2016.
- [30] Ioffe, Sergey, and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proceedings of the International Conference on Machine Learning. 2015.
- [31] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. International Conference on Learning Representations. 2015.