big data and cognitive computing

MDPI

Article

Analyzing the Attractiveness of Food Images Using an Ensemble of Deep Learning Models Trained via Social Media Images

Tanyaboon Morinaga [1], Karn Patanukhom [2,3,*] and Yuthapong Somchit [2]

[1] Data Science Consortium, Faculty of Engineering, Chiang Mai University, Chiang Mai 50200, Thailand; tanyaboon_m@cmu.ac.th
[2] Department of Computer Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai 50200, Thailand; yuthapong@eng.cmu.ac.th
[3] Advanced Technology and Innovation Management for Creative Economy Research Group, Chiang Mai University, Chiang Mai 50200, Thailand
* Correspondence: karn.patanukhom@cmu.ac.th

Abstract: With the growth of digital media and social networks, sharing visual content has become common in people's daily lives. In the food industry, visually appealing food images can attract attention, drive engagement, and influence consumer behavior. Therefore, it is crucial for businesses to understand what constitutes attractive food images. Assessing the attractiveness of food images poses significant challenges due to the lack of large labeled datasets that align with diverse public preferences. Additionally, it is challenging for computer assessments to approach human judgment in evaluating aesthetic quality. This paper presents a novel framework that circumvents the need for explicit human annotation by leveraging user engagement data that are readily available on social media platforms. We propose procedures to collect, filter, and automatically label the attractiveness classes of food images based on their user engagement levels. The data gathered from social media are used to create predictive models for category-specific attractiveness assessments. Our experiments across five food categories demonstrate the efficiency of our approach. The experimental results show that our proposed user-engagement-based attractiveness class labeling achieves a high consistency of 97.2% compared to human judgments obtained through A/B testing. Separate attractiveness assessment models were created for each food category using convolutional neural networks (CNNs). When analyzing unseen food images, our models achieve a consistency of 76.0% compared to human judgments. The experimental results suggest that the food image dataset collected from social networks, using the proposed framework, can be successfully utilized for learning food attractiveness assessment models.

Keywords: CNNs; image aesthetic quality assessment; social media; food image; image classification; attractiveness

Citation: Morinaga, T.; Patanukhom, K.; Somchit, Y. Analyzing the Attractiveness of Food Images Using an Ensemble of Deep Learning Models Trained via Social Media Images. *Big Data Cogn. Comput.* **2024**, *8*, 54. https://doi.org/10.3390/bdcc8060054

Academic Editors: Robail Yasrab and Md Mostafa Kamal Sarker

Received: 18 March 2024
Revised: 21 May 2024
Accepted: 22 May 2024
Published: 27 May 2024

1. Introduction

In the era of digital media and social networks, the proliferation of visual content has been unprecedented. A massive number of images are shared daily on platforms like Instagram, Facebook, and X (formerly Twitter), capturing various aspects of our lives, from personal moments to artistic expressions. One domain that has witnessed a surge in visual content sharing is the restaurant and food industry. Restaurants increasingly rely on visually appealing images of their dishes and ambiance to attract customers and promote their brand on social media [1], where people share and communicate through images. One key aspect of food content on social media is its attractiveness, which is often determined by the visual appeal of the food images. Attractive food images can attract attention, engagement, and even drive sales for businesses that rely on food as a product or service [2]. As a result, it is important for businesses and individuals to understand what makes a food image attractive to their audience.

Currently, convolutional neural networks (CNNs) are state-of-the-art in many image classification tasks, showing remarkable success in various fields, such as facial recognition [3,4], medical applications [5–7], or agricultural applications [8–10]. A CNN is a type of artificial neural network that utilizes convolutional layers in the model. The convolutional layers result from the convolution of the input with various kernels. Applying convolutional layers to neural networks can reduce the number of trainable parameters in comparison with fully connected layers since convolutional layers allow multiple weights to share values. For image recognition tasks, the two-dimensional convolution operator allows neural networks to learn local patterns such as edges, textures, or shapes that appear in the input data. Cascading multiple convolutional layers and down-sampling layers makes neural networks capture more complex patterns of shapes or textures. For food images, CNNs can be utilized for food category or menu recognition [11–13], calorie estimation [14,15], or attractiveness assessments [16,17]. In supervised learning scenarios, where CNNs excel, the model's accuracy heavily depends on the volume and quality of its training data. Although there are many labeled datasets for food category or menu recognition tasks, such as ISIA Food-500 [18], which contains 399,726 images from 500 classes of foods, Food2K [19], which contains more than 1 million images from 2000 classes of foods, or FruitVeg-81 [20], which contains 15,737 images from 81 classes of vegetables and fruits, there are few datasets that can be used for food attractiveness assessments. The reason for this is that the preparation of such datasets for assessing food image attractiveness can be labor-intensive and challenging. It requires not only collecting a vast number of food images but also ensuring that these images are labeled accurately. More importantly, the labels need to align with the diverse and subjective preferences of the general public rather than relying solely on the opinions of a few (expert) annotators, like menu classification tasks.

This research presents a novel data-driven approach that circumvents the need for explicit human annotation by leveraging user engagement data that are readily available on social media platforms. It demonstrates that food images appearing on social media platforms, along with their corresponding user engagement metrics such as the number of likes, can be applied to train models to evaluate their attractiveness. However, appropriate data processing is necessary since the level of user engagement can be influenced by factors other than the attractiveness of the image itself, particularly the poster's identity, the duration the post has been up, or the polarity of engagement (positive or negative sentiments).

This work presents a novel method for preparing food image datasets and using CNN training, bypassing the need for manual labeling by utilizing poster engagement data from social media as indicators of the visual attractiveness of the images. We also propose procedures to collect images as training data, automatically label attractiveness classes for the training data, and create an attractiveness assessment model. We also proposed methods to mitigate the previously mentioned bias factors. Regarding the popularity of posters that can affect the number of engagements, we propose separately ranking the engagement of images among individual posters and developing multiple predictive models based on each poster's dataset independently. Additionally, for the time of posting and duration since posting, which might impact the number of engagements, the recently posted images are removed from the training data to ensure that the level of user engagement for each image is stable and reflective of its actual attractiveness level. To simplify the problem, in this work, we first focus on studying the category-specific attractiveness assessment models and test them on five food categories, including sushi, ramen, pizza, burger, and cake.

To summarize the contributions of this work, they are as follows:

1. We propose a novel end-to-end framework for analyzing the attractiveness of food images, which includes data collection, filtering, automatic labeling, and predictive ensemble model development.
2. We conduct experiments to demonstrate that the food image dataset collected from social networks can be automatically annotated and utilized to develop the attractiveness assessment model in a supervised manner.

3. We conduct experiments using Grad-CAM to visualize the important regions in the food images that affect the attractiveness score.

The paper is organized as follows: Section 2 will present some existing research related to this work, Section 3 will present the proposed food attractiveness assessment methodology, Section 4 will present our experiment and results, and the conclusion of this work will be presented in Section 5.

2. Literature Review

In general, there are two perspectives in image assessment. The first one, image quality assessment (IQA) [21–23], focuses on the perceptual quality affected by factors such as noise, distortion, blur, compression, etc. The other is image aesthetic quality assessment (IAQA) [24–29], which measures image quality in artistic aspects affected by lighting conditions, color settings, camera direction, image composition, etc. Human judgments of aesthetic quality depend on their experiences and may differ among individuals. The techniques for computer-based IAQA can be categorized into two groups: handcrafted feature-based methods [24,25] and deep learning-based methods [26,29]. Our work can be considered a type of IAQA, which we call food image attractiveness assessment.

Currently, numerous studies have presented guidelines for image processing, focusing on aesthetics and attractiveness assessments. These studies involve creating training datasets to evaluate aesthetics and testing them with various machine learning algorithms and visual features. It has been concluded that these methods can perform effectively. One notable work is the study by Sheng et al. [30], who introduce the Gourmet Photography Dataset (GPD). This large-scale dataset is specifically designed for assessing the aesthetics of food images. The GPD contains 12,000 food images aesthetically scored by human annotators on Amazon Mechanical Turk. It serves as a benchmark for training and evaluating models for food image aesthetic assessment and has been utilized in various challenges. In this study, various machine learning models are tested on the GPD. CNN-based models can achieve testing accuracy in the range of 77.25–90.79%.

Another work by Sheng et al. [31] introduces a new regularization method to improve model generalization on aesthetic assessment tasks. This work elaborates on the difficulties of distinguishing aesthetically pleasing food images and the limitations of the datasets available for this purpose. The GPD presented here contains 24,000 images, an increase from the initial offering, emphasizing binary aesthetic labels across a broad range of food types and scenes. A novel contribution of this paper is the introduction of a non-stationary regularization method, adaptive smoothing regularization (ASR), which is designed to combat overfitting and enhance model generalization. The paper demonstrates that neural networks trained with ASR on the GPD can achieve comparable performance to human experts, offering insights and support for further research in visual aesthetic analyses of food images.

The work of Takahashi et al. [16,17] presents methods for estimating food image attractiveness based on the visual characteristics of main ingredients and image feature analysis. Their innovative approach combines analysis of the overall food image impression with a detailed evaluation of the main ingredient, enriching the understanding of aesthetic appeal in food images, which is particularly relevant in social media and digital marketing contexts. Their method involves extracting image features that focus on both the appearance of the entire food item and its main ingredients. Attractiveness is then estimated through a regression scheme that integrates these features. A specially constructed and publicly released food image dataset, containing images of ten food categories taken from 36 angles with accompanying attractiveness values, was used for evaluation. The results demonstrated the effectiveness of integrating two kinds of image features for estimating the attractiveness of food images. The quantitative outcomes of their research are compelling. The average mean absolute error (MAE) of the proposed method was 0.087, significantly outperforming the comparative method, with an MAE of 0.344. This indicates superior accuracy in estimating food image attractiveness by considering specific food

image characteristics rather than general aesthetic qualities. The research suggests that the attractiveness of food photography can be quantified and predicted with a high degree of accuracy using the proposed method. This capability can support systems that recommend the best camera framing for capturing attractive food images or assist in selecting the most appealing image from a set, ultimately enhancing the visual appeal of food images shared on social media platforms.

Philp et al. [32] examine the relationship between the visual characteristics of food images on Instagram and social media engagement. They utilize Google Vision AI, an image classification machine learning algorithm, to analyze the confidence of food image classification and use it as a proxy for the visual typicality of food images. Their findings suggest that images deemed more visually typical are positively associated with higher social media engagement. The quantitative analysis reveals a positive correlation: as the visual typicality of food images increases, so does engagement, as measured through likes and comments. Specifically, the study highlights that for each unit increase in food typicality, there is a significant uptick in the number of both likes and comments. This challenges the prevailing notion in marketing that uniqueness drives engagement, suggesting instead that familiarity may be equally, if not more, compelling in attracting audience interaction on social platforms.

The work of Attokaren et al. [33] presents a comprehensive study on food classification CNNs, highlighting the importance of accurate food monitoring for health reasons. The authors discuss the challenges in food image classification, the potential of CNNs to overcome these challenges, and the applications of this technology in health monitoring and dietary management. Utilizing the Food-101 dataset and a pre-trained InceptionV3 model [34], they achieve an accuracy of 86.97% in food image classification. This research not only demonstrates the effectiveness of CNNs in classifying food images but also emphasizes their importance for applications in smart dietary management technologies. The high accuracy achieved underscores the potential of leveraging CNNs for enhancing food recognition systems, thereby contributing to advanced solutions in health monitoring and nutritional advice.

Islam et al. [35] develop a CNN model specifically designed to classify food images. Their work aims to address the significant intra-class variability found within images of the same food category, a major challenge in image classification. Utilizing the Food-11 dataset and employing the pre-trained InceptionV3 model, they demonstrate the potential of CNNs to effectively extract spatial features from food images. This research is particularly significant for social media platforms and restaurants, providing a tool for identifying and categorizing food images, which can be beneficial for targeted advertising and enhancing user engagement. The experimental results were impressive, with the pre-trained InceptionV3 model achieving an accuracy of 92.86%, significantly outperforming the scratch-built CNN model, which achieved an accuracy of 74.70%. This highlights the effectiveness of using pre-trained models for food image classification in overcoming the challenges of significant intra-class variability within food categories.

From the related work presented in this section, we can observe the high potential of CNN models for analyzing food images. Unfortunately, there are few studies related to the assessment of attractiveness in food images. To the best of our knowledge, there is no work that has proposed a procedure to utilize social media data without real attractiveness annotations to train an attractiveness assessment model.

3. Proposed Methods

3.1. Overview

Our proposed methods consist of two phases: the learning phase and the inference phase. The illustrations of each phase are shown in Figures 1 and 2. In the learning phase, we collect data from multiple posters within the target food categories. Subsequently, data cleaning and automatic labeling processes are applied to the dataset. Finally, multiple models are trained using data from each poster to train individual models. As a result,

if data are collected from N posters, there will be N models for the assessment of attractiveness. In the inference phase, the target unseen image is input to each model, resulting in N predictions. Then, a voting ensemble is utilized to combine the results from the N models, producing the final prediction of the attractiveness class. The details of each process will be described in the following sections.

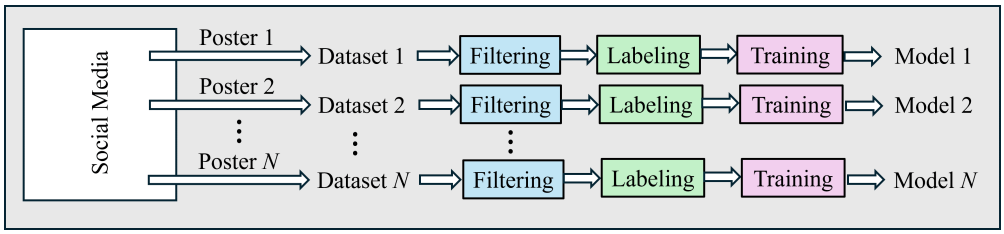

Figure 1. Overview of the learning phrase of the proposed attractiveness assessment model.

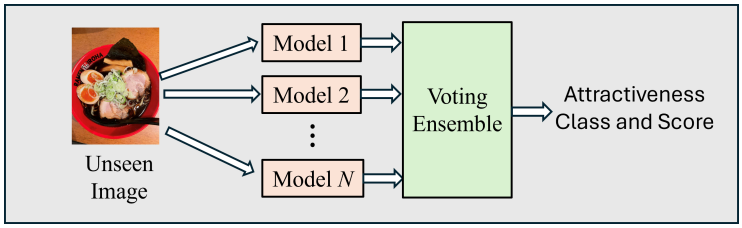

Figure 2. Overview of the inference phrase of the proposed attractiveness assessment model.

3.2. Data Collection

In the data collection process, we have to search for user accounts of potential food image posters for the target food category on a specific social media platform. The process starts by searching for public images using the target food category name (e.g., sushi, ramen, pizza) as a keyword. Poster accounts are collected from the search results for further investigation. Next, the images posted on each account are inspected to determine whether the poster primarily focuses on a target food category. Data collection efforts focus on accounts dedicated to reviewing and showcasing specific types of food, as these accounts tend to feature a concentrated collection of images within a particular food category. If a poster consistently posts images of a target food category, and the number of relevant images exceeds the threshold M_{min}, the account is considered suitable for inclusion in our data collection process. The value of M_{min} can vary depending on the social media platforms and the target food categories. Increasing the value of M_{min} can improve the data quantity for individual model training but may decrease the number of models used in the voting ensemble process as the number of qualified posters decreases. Since each dataset from a single poster will be used to train one model, the dataset must only contain images of a single food category that meet the minimum number threshold (M_{min}). For each qualified poster, we create a dataset that contains the user profiles, the set of posted images, and their corresponding data, such as the number of likes and comments and the posted date.

3.3. Data Filtering

The initial step in preprocessing involves filtering out irrelevant images:

- Duplicate removal : Any duplicate images are identified using hashing techniques and subsequently removed to ensure the uniqueness of each image in the dataset. To ensure that there are no duplicated images, a manual check is performed afterward.
- Irrelevant image exclusion: Images that do not relate to food in the target category, such as people, scenery, pets, or images from other food categories, are excluded in this step. This ensures that the dataset strictly focuses on the target food images. In this paper, a manual filtering process is employed to achieve this exclusion; however, it can also be performed automatically using machine-learning models [33,36–39].

- Recency filter: Normally, newly posted images tend to initially have lower engagement levels, but the level of engagement increases rapidly in the early stages. The rate of increase gradually decreases over time, and once a sufficient amount of time has passed, the number of engagements tends to stabilize or change minimally. To ensure stability in user engagement metrics, recently posted images were excluded. The optimal timing threshold T_{min} can be obtained based on the knee point of the user engagement curve over the posting duration.

3.4. Automatic Attractiveness Class Labeling

Once the dataset is cleaned, the next step is labeling. As mentioned before, the level of user engagement cannot be directly used as the level of attractiveness since it may be influenced by the duration that the post has been up and the poster's identity. By using the recency filter in the previous step, we can now assume that the level of user engagement in the cleaned data is independent of the duration that the post has been up. Additionally, the dataset for training each model for each food type is collected from separate posters. Each poster may have a different number of followers, which influences the user engagement level on the images compared to other posters. Using the dataset separately for each model can reduce the bias associated with using the user engagement level as a measure of attractiveness. For the same poster, more attractive images typically receive more user engagement than less attractive ones. As a result, the level of user engagement for training each model is independent of the poster's identity and can serve as a potential representation of the attractiveness level.

To train a model to learn the differences in attractiveness levels, we sort the data based on their user engagement levels. The images with the top-k percentile user engagement are labeled as in the highly attractive (H) class while, the images with the bottom-k percentile user engagement are labeled as in the less attractive (L) class. Now, we have obtained the labeled data balanced between two classes.

The unlabeled data between the top-k percentile and the bottom-k percentile are considered the blurred boundary between the distributions of two attractiveness classes and will not be used for model training. The value of k must be less than 50%. Decreasing the value of k results in a training dataset with higher confidence in automatic class labeling, though the number of samples will decrease. This represents a trade-off between the quality and quantity of the dataset.

3.5. Attractiveness Assessment Model

In this work, we simplify the attractiveness assessment problem into the binary classification problem of H and L classes. In this work, we utilize CNNs as the classifier because they are currently state-of-the-art techniques. However, to implement this in the real world, the classifier can be extended to use other machine learning techniques as well as CNNs. Any pre-trained CNN-based classifier, such as VGG [40], ResNet [41], Inception [34], MobileNet [42], or EfficientNet [43], can be utilized as the backbone of our attractiveness assessment model. In the training phase, N binary classifiers are trained separately using a labeled dataset collected from each qualified poster. Using a CNN model, each classifier can learn the difference between highly and less attractive images via visual features. Training multiple models from different sources may help each individual model learn attractiveness from different aspects. After all models are trained, we can use them to predict the attractiveness class of any image. In the inference phrase, the target image is input to each classifier, with the output being the prediction of class H or L. In the final step, hard voting is applied to all predicted results to obtain the final attractiveness class. In addition, we can calculate the attractiveness score $\in [0, 1]$ as $\frac{N_H}{N}$, where N_H is the number of models that predict the target image as class H.

4. Experiment and Results

4.1. Data Collection

In the experiment, we evaluated the performance of our proposed food attractiveness assessment procedure on five food categories, which are sushi, ramen, pizza, burger, and cake. These categories were chosen because they are representative of popular food types, with a large number of posters meeting our criteria. Additionally, they provide a diverse range of visual styles and aesthetics, making them suitable for testing our approach to assessing the visual appeal of food images. We chose Instagram as the social media platform to collect the data because it is the social media platform where users primarily communicate by posting images. There are many accounts that specialize in food reviews of the target categories, ensuring a diverse and representative set of samples.

The data were collected, filtered, and labeled using the process described in Sections 3.2–3.4. In this experiment, we used $M_{min} = 1000$ for sushi, ramen, burger, and cake, and $M_{min} = 800$ for pizza because the number of pizza images posted is less than the other categories.

The process of collecting the data is shown in Figure 3. We scraped data from Instagram to find posters with the number of images of one food type that satisfy the above conditions. One food type required 10 posters. We found a total of 50 posters with a total of 168,361 images of all five food types. Then, the images were processed by data filtering, and 112,190 (66.6%) remained after the process. Afterward, the data were divided into two sets: the development set and the A/B testing set.

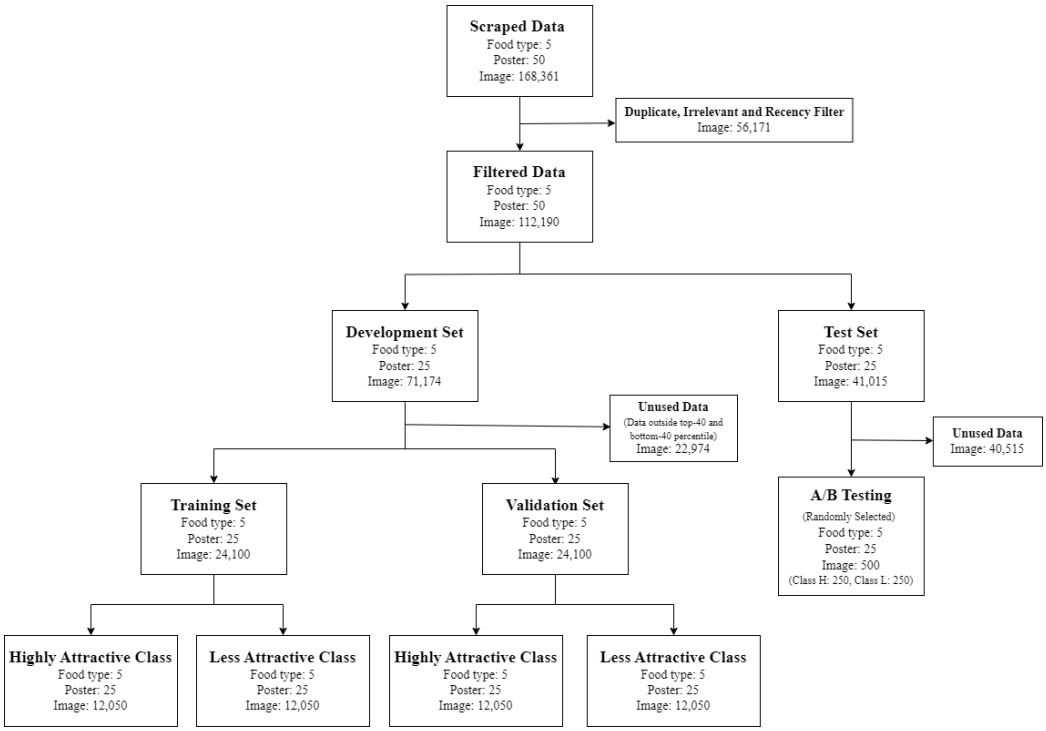

Figure 3. Data collection and processing flow diagram.

For the development set, five posters per food type were randomly selected to train five individual classifiers, i.e., one poster for one model. As a result, there are five models for one food type, with a total of twenty-five models for all five food types. In this experiment, we used $k = 40\%$ for automatic attractiveness class labeling, which means images within the top 40th percentile were tagged as class H and images within the bottom 40th percentile were tagged as class L. There are a total of 22,074 images in the development set that were not tagged into any class and are unused in this experiment. Finally, the development set was partitioned into a training set and a validation set with a ratio of 1:1, resulting in 24,100 images for the training set and 24,100 images for the validation set. Half of the images are class H, and the other half are class L. The number of images for training each

individual model in each food category is demonstrated in Table 1. For the A/B testing set, we randomly chose 10 pairs of class H and class L images from each poster, resulting in a total of 250 H-class images and 250 L-class images.

Table 1. Size of training sets.

Model	Sushi	Ramen	Pizza	Burger	Cake
Model 1	1400	800	1500	1200	1100
Model 2	1300	700	1120	1500	800
Model 3	600	660	600	800	600
Model 4	800	1800	600	740	900
Model 5	1800	900	400	680	800
Total	5900	4860	4220	4920	4200

In this experiment, we used the number of likes as the positive user engagement metric, which was converted into the attractiveness class. We examined the curve of the number of likes over the posting duration. Figure 4 shows the relationship between the average number of likes and the posting duration (gray line) and the samples of the number of likes for each image from three posters (color dots). We can observe that the number of likes changes rapidly over the first five days (knee point) and slows down after seven days. To ensure stability in the number of likes, images posted less than seven days ago were excluded ($T_{min} = 7$). By removing these images, we can ensure that the number of likes for each image in our dataset is stable and reflective of its actual attractiveness level.

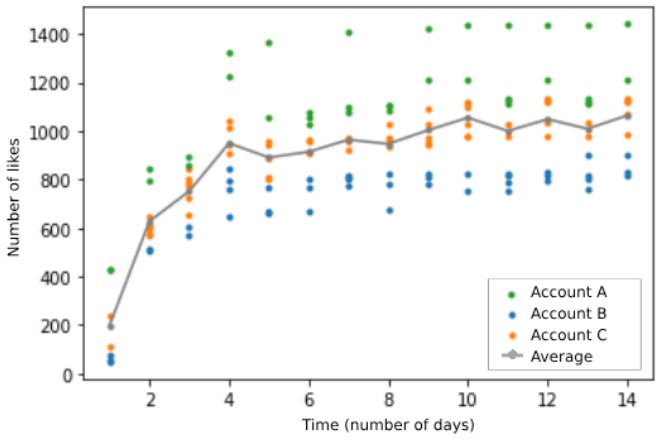

Figure 4. Number of image likes, which increased since the image was posted.

We conducted further investigations to confirm that the time and date of posting do not impact the number of likes. The number of likes from images across our dataset was plotted against the posting time and day, as shown in Figures 5 and 6, respectively. The results consistently demonstrate a similar range in the number of likes, regardless of what time or on what day the posts were made. These findings verify that the time and day of posting do not affect the number of likes in our datasets.

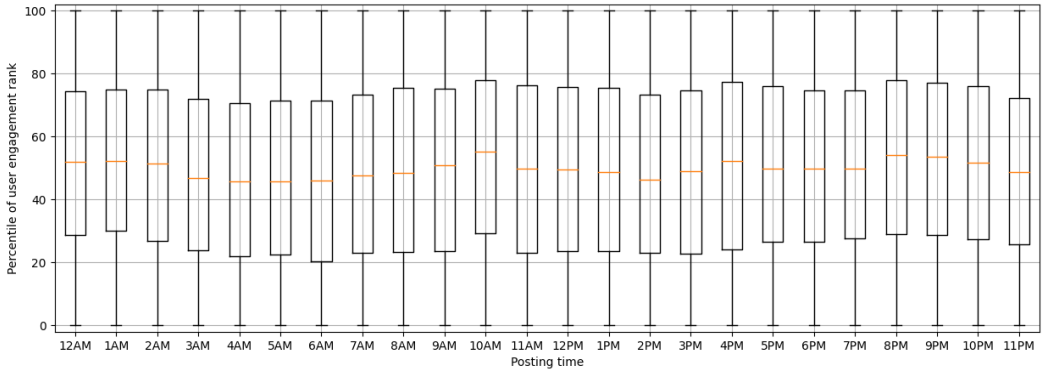

Figure 5. Boxplot of the number of likes against the time of posting.

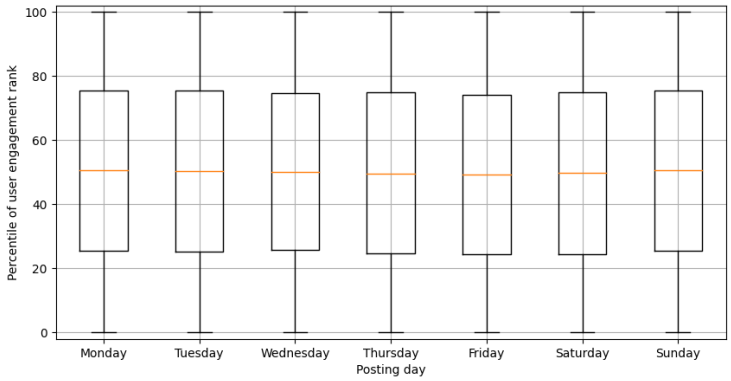

Figure 6. Boxplot of the number of likes against the day of posting.

4.2. A/B Testing

To assess the effectiveness of our model, we conducted A/B testing. For each food category, we selected 50 pairs (H-class image, L-class image) from the testing set (images from different posters than those in the training set). We gathered 51 participants to perform the A/B testing. The participants were blind to the actual like counts of the images. During A/B testing, 50 pairs of images were shown to the participants, with the order of images in each pair randomized, as shown in Figure 7. Participants were asked to select the image they found most attractive in each pair. When we compared the A/B testing results with our proposed automatic class annotation utilizing the number of likes, there was a 97.2% consistency between the A/B testing results and the automatic class annotation results. The consistency percentage is defined as the ratio of the number of images in class H, which more than 50% of the participants selected as more attractive in A/B testing, to the number of testing pairs. The detailed results in each category are shown in Table 2. The burger category has the highest consistency of 100%, while the sushi category has the lowest consistency of 94%. The results of the A/B testing show that we successfully mapped the number of likes to the real attractiveness level.

Table 2. The consistency percentage between the A/B testing results and the automatic class annotation results.

Food Category	Sushi	Ramen	Pizza	Burger	Cake
Consistency	94%	96%	98%	100%	98%

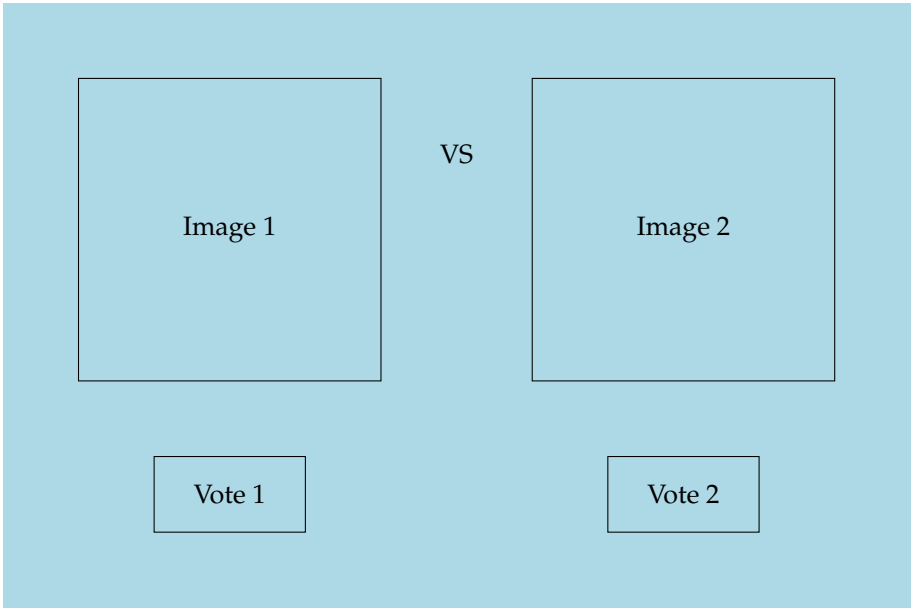

Figure 7. Interface of A/B testing for image attractiveness.

Table 3 shows how variations in the values of the k parameter impact the consistency of the A/B testing results and the training set size. From these results, it can be concluded that varying the k parameter does not affect consistency. However, larger values of k result in a larger training dataset size. Therefore, we used $k = 40\%$ in this experiment.

Table 3. The variations in the values of the k parameter with the consistency of A/B testing results and the training set.

Parameter k	40%	30%	20%	10%
Consistency	97.2%	97.4%	97.8%	98.1%
Training Set Size	24,100	18,075	12,050	6025

4.3. Model Development and Training

In this experiment, we selected InceptionV3 [34], pre-trained on the ImageNet dataset [44], as the backbone of our attractiveness classification model. We extended this architecture by appending a global average pooling 2D layer to reduce the dimensions of the feature maps and minimize overfitting. This is followed by a dropout layer set at a rate of 0.2 to further regularize the model during training. The final layer is a dense layer with a softmax activation function that outputs the probability distribution over the two classes. The number of convolutional layers is 94, and the number of trainable parameters is 21,772,450. The architecture of InceptionV3 is shown in Figure 8. These models are implemented using the TensorFlow library.

Image augmentation is applied during the learning phase. The augmentation parameters include random transformations like shearing, zooming, and horizontal flipping to improve the model's ability to generalize from our dataset. All models use the stochastic gradient descent (SGD) optimizer with the same learning rate of 0.0001 and momentum of 0.9. The loss function used is categorical cross-entropy. Training is conducted over 5000 epochs, a large number, which is intended to ensure thorough learning, with early stopping implemented to prevent overfitting. During training, the model's accuracy on the validation set is monitored, and only the best-performing weights are saved.

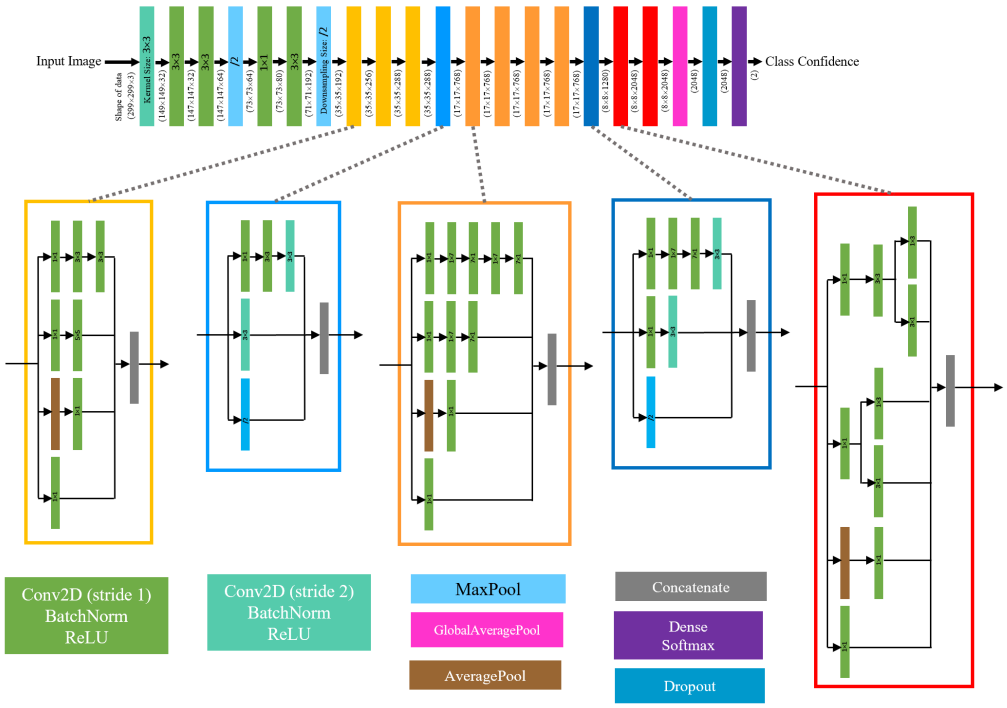

Figure 8. Architecture of the InceptionV3 model.

The training and validation accuracies are reported in Tables 4 and 5. Models 1 through 5, as shown in the tables, are classifiers that were trained and validated using datasets collected from different posters. Therefore, any model *i* represented across different food categories in the tables refers to distinct models. The results indicate that training accuracies vary between 99% and 100%, while validation accuracies range from 55% to 99%.

Table 4. Training accuracy of all models across each food category.

Model	Sushi	Ramen	Pizza	Burger	Cake
Model 1	100%	100%	100%	100%	100%
Model 2	99%	100%	100%	100%	100%
Model 3	100%	100%	99%	100%	100%
Model 4	100%	100%	99%	100%	100%
Model 5	100%	100%	100%	100%	100%

Table 5. Validation accuracy of all models across each food category. The best validation accuracy of each category is highlighted in yellow.

Model	Sushi	Ramen	Pizza	Burger	Cake
Model 1	99%	79%	69%	81%	90%
Model 2	73%	66%	98%	70%	88%
Model 3	78%	75%	55%	90%	75%
Model 4	75%	68%	74%	89%	81%
Model 5	80%	73%	75%	74%	71%

4.4. Model Performance Evaluation

4.4.1. Quantitative Evaluation

The proposed attractiveness assessment models were evaluated on 100 images for each food category, corresponding to 50 pairs in A/B testing, totaling 500 images. Similar to human evaluators in A/B testing, the models were tasked with selecting the more attractive image in each pair by comparing the predicted attractiveness scores. For the individual model, the attractiveness score is equal to the predicted probability of class H. For the ensemble model, the attractiveness score is calculated from the ratio of class H predictions made by the individual models to the total number of models.

The A/B testing results from the models are compared with (1) automatically labeled classes based on the number of likes and (2) the voting results of the human participants in A/B testing. The results are presented in Tables 6 and 7, respectively. Yellow backgrounds indicate the best individual model results for each food category, while blue backgrounds indicate the ensemble results.

Table 6 presents the performance of various models in predicting the visual appeal of food images across five categories: sushi, ramen, pizza, burger, and cake. The predictions are compared against automatically labeled attractiveness classes. In this table, the consistency is calculated from the ratio of the number of images in class H that the model selects as more attractive in A/B testing to the number of testing pairs. For ensemble models, the best consistency of 90% is obtained in the sushi category, while the worst consistency of 66% is obtained in the cake category. The information from Tables 1 and 6 shows that the prediction performance tends to increase as the number of training samples increases.

Next, in Table 7, we compared the models' judgments in A/B testing with the human participants' judgments. In this table, the consistency is calculated as the ratio of the number of pairs where the model and the majority of the human participants select the same image as the more attractive image to the number of testing pairs. The results indicate that the ensemble models demonstrate consistent alignment with human judgments, with rates of 84% for sushi, 76% for ramen, 72% for pizza, 80% for burger, and 68% for cake images.

Table 6. Consistency of the individual and ensemble model predictions in comparison to automatically labeled attractiveness classes. The best consistencies obtained from the individual models are highlighted in yellow, while those from the ensemble models are highlighted in blue.

Model	Sushi	Ramen	Pizza	Burger	Cake
Model 1	74%	74%	82%	80%	56%
Model 2	70%	66%	66%	66%	58%
Model 3	82%	48%	60%	88%	68%
Model 4	76%	76%	72%	72%	74%
Model 5	88%	68%	52%	58%	50%
Ensemble	90%	76%	74%	80%	66%

Table 7. Consistency of the individual and ensemble model prediction in comparison to the human voting results of A/B testing. The best consistencies obtained from the individual models are highlighted in yellow, while those from the ensemble models are highlighted in blue.

Model	Sushi	Ramen	Pizza	Burger	Cake
Model 1	72%	70%	80%	80%	54%
Model 2	68%	70%	64%	66%	60%
Model 3	80%	44%	62%	88%	66%
Model 4	70%	80%	74%	72%	76%
Model 5	82%	68%	50%	58%	52%
Ensemble	84%	76%	72%	80%	68%

The results in Table 6 align with those obtained in Table 7 because of the high consistency between the automatic class labeling and A/B testing results from human voters, as previously shown in Table 2. For burger, the results in Tables 6 and 7 are exactly the same. The most significant difference in the results between Tables 6 and 7 is sushi, which has a consistency with A/B testing results that is 6% lower than the consistency with the automatic class labeling. Four out of the five food categories have ensemble consistencies lower than those of the best individual models. The reason for this is that the performance of each individual model in the ensemble process is diverse. For example, in the case of pizza, the consistency of each individual model in Table 7 varies from 50%, 62%, 64%, 74%, to 80%. Theoretically, if we assume that all individual models are independent, the expected consistency obtained from the ensemble model is 78.6%. The experimental result has a consistency of 72%, which is 6.6% lower than the expected result. On the other hand, for sushi, the consistency of each individual model listed in Table 7 exhibits less variation (68%, 70%, 72%, 80%, and 82%) and a higher mean compared to the previous example. Theoretically, the expectation of the consistency obtained from the ensemble model is 89.2%. In this case, the expectation of ensemble consistency is higher than that of all individual consistencies. The experiment achieved a consistency of 84%, which is 5.2% lower than the expectation.

It is worth noting that some variations in performance exist across different cuisine categories, which may be attributable to factors such as the diversity of visual styles, cultural influences, or the specific characteristics of the training data within each category. However, the overall results are encouraging and underscore the promise of our approach in developing scalable and efficient solutions for computational aesthetic assessment, particularly in the domain of food images.

Moreover, we conduct a fine-grained evaluation to observe the consistency between the model-generated attractiveness scores and the human voting results in A/B testing. Let P be the image that receives more votes from human participants, and let N be the image that receives fewer votes. We calculate scores for comparison based on the differences in the number of votes between images P and N from both the human participants and the ensemble model, as detailed in Equations (1) and (2).

$$\Delta score_H = \frac{H_P - H_N}{H_P + H_N} \tag{1}$$

$$\Delta score_M = \frac{M_P - M_N}{M_P + M_N} \tag{2}$$

In the equations, H_P and M_P represent the number of votes for image P in the A/B testing from the human participants and the ensemble model, respectively. Similarly, H_N and M_N represent the number of votes for image N. $\Delta score$ shows the difference in attractiveness for each testing pair in the A/B testing. The terms $\Delta score_H$ and $\Delta score_M$ indicate the score differences obtained from the human participants and the ensemble model, respectively. We categorized the values of $\Delta score$ into six classes, referred to as Δ-class, as shown in Table 8. The consistency of the Δ-class obtained from the model and the human results is illustrated as confusion matrices in Tables 9 and 10. Table 9 shows results from ensemble models. Table 10 shows results from the best individual models. In these tables, a green background indicates that the votes from the model and humans are exactly the same. Light green signifies that votes from the model and humans differ by one or two levels of the Δ-class. A red background indicates that the winner of attractiveness voted by the model differs from that voted by the humans.

It can be concluded from the results in Table 9 across all five food categories that 25.6% of the samples have a $\Delta score$ from the model voting in the same Δ-class as from the human voting, 66.8% are within one level of Δ-class difference, and 87.6% are within two levels of Δ-class difference. On the other hand, for the best individual model in Table 10, 20.0% of the samples have a $\Delta score$ from the model voting in the same Δ-class as from the human voting, 60.0% are within one level of Δ-class difference, and 89.2% are within two levels

of Δ-class difference. The fine-grained results from the ensemble model are closer to the human votes than those from the best individual models.

Additionally, considering the attractive vote winner that differs between the model and humans, there are a total of 60 samples where the ensemble model votes differ from those of the humans in the A/B testing. We found that in 71.6% of these cases, the Δ-class is F1, which indicates that the ensemble models selected most of the incorrect images in the A/B testing with low confidence.

Table 8. Definition of Δ-class.

Δ-Class	Range of Δ*Score*	Definition
T3	$(0.8, 1.0]$	Image P is significantly more attractive than image N.
T2	$(0.4, 0.8]$	Image P is moderately more attractive than image N.
T1	$(0.0, 0.4]$	Image P is slightly more attractive than image N.
F1	$(-0.4, 0.0]$	Image N is slightly more attractive than image P.
F2	$(-0.8, -0.4]$	Image N is moderately more attractive than image P.
F3	$[-1.0, -0.8]$	Image N is significantly more attractive than image P.

Table 9. Confusion matrix of ensemble model in A/B testing.

		Sushi			**Ramen**			**Pizza**			**Burger**			**Cake**		
		$\Delta Score_H$			$\Delta Score_H$			$\Delta Score_H$			$\Delta Score_H$			$\Delta Score_H$		
		T3	T2	T1	T3	T2	T1	T3	T2	T1	T3	T2	T1	T3	T2	T1
$\Delta score_M$	T3	9	9	0	6	3	0	5	3	1	9	4	1	3	4	0
	T2	10	4	0	7	4	2	4	10	2	12	4	1	7	4	0
	T1	2	6	2	11	3	2	1	9	1	4	4	1	11	5	0
	F1	1	1	1	4	4	2	2	4	4	5	2	1	6	6	0
	F2	0	2	2	1	0	0	0	2	1	0	1	0	1	1	1
	F3	0	0	1	0	0	1	1	0	0	0	0	1	1	0	0

Table 10. Confusion matrix of the best individual model in A/B testing.

		Sushi			**Ramen**			**Pizza**			**Burger**			**Cake**		
		$\Delta Score_H$			$\Delta Score_H$			$\Delta Score_H$			$\Delta Score_H$			$\Delta Score_H$		
		T3	T2	T1	T3	T2	T1	T3	T2	T1	T3	T2	T1	T3	T2	T1
$\Delta score_M$	T3	5	2	0	6	6	1	4	11	5	9	6	2	3	6	0
	T2	2	2	0	6	4	0	2	3	0	8	3	1	8	2	0
	T1	14	13	3	12	2	3	5	8	2	10	4	1	10	9	0
	F1	1	4	3	3	1	1	2	4	2	3	1	0	6	2	0
	F2	0	1	0	1	1	2	0	2	0	0	1	0	0	1	0
	F3	0	0	0	1	0	0	0	0	0	0	0	1	2	0	1

These results show the efficiency of our approach in leveraging Instagram to increase the number of likes and create labeled datasets for training aesthetic assessment models. The high degree of alignment between the models' predictions and human perceptions highlights the potential of our method to capture and encode visual appeal accurately without the need for labor-intensive manual annotation.

4.4.2. Qualitative Evaluation

To analyze the factors that play important roles in image attractiveness, 32 sample images from the proposed framework's attractiveness grading results are illustrated in Figure 9. Food images within the same category, which are quite similar to each other, are grouped together, resulting in 12 groups for use in comparison. Color settings and lighting intensity impact attractiveness, as demonstrated in the Ramen-A (samples 8 and 9), Ramen-C (samples 13 and 14), and Pizza-A (samples 15, 16, and 17) groups. A warmer tone of lighting is shown to reduce attractiveness, as seen in the Ramen-A group, where sample 9 has a lower attractiveness score than sample 8, and in the Ramen-C group, where sample 14 scores lower than sample 13. Additionally, lighting temperature affects attractiveness scores; overly bright or dark lighting reduces the attractiveness score. This is demonstrated in the Pizza-A group, where sample 17 has a higher attractiveness score than samples 15 and 16. The number of foods in the image also affects attractiveness scores. This is evident in the Sushi-A (samples 1, 2, and 3) and Pizza-B (samples 18, 19, and 20) groups, where focusing on a single piece of sushi or pizza yields a higher attractiveness score. Similarly, results from the Cake-B group (samples 31 and 32) show that images containing both a drink and a cake have lower attractiveness scores compared to those focusing solely on a cake. In the case of burgers, results from the Burger-A (samples 21, 22, 23, and 24) and Burger-B (samples 25, 26, and 27) groups demonstrate that presenting more meats in the images can lead to higher attractiveness scores.

For an in-depth analysis of the results, gradient-weighted class activation mapping (Grad-CAM) [45] is used. Grad-CAM is a technique that produces class activation maps highlighting important regions in an image for a classifier to predict each class probability. Let G_H and G_L represent the activation maps for classes H and L, respectively. Figure 10 shows an example of Grad-CAM analysis from five models for one image. G_H highlights the regions that positively influence attractiveness, while G_L highlights regions that negatively influence attractiveness. Since each model learns the differences between two classes of attractiveness from different datasets, the important regions identified by each model may vary. For instance, consider the sliced meat region in the ramen image shown in Figure 10. In models 4 and 5, this region positively influences attractiveness, whereas in model 1, a similar region negatively influences attractiveness.

Figure 11 displays examples of G_H from fifteen food images extracted from five models, revealing patterns in their focus areas. In the case of Sushi 1, all models focus on the same area, demonstrating high confidence in the prediction. For Sushi 2, four out of five models focus on the salmon sushi, which is the most colorful piece in the image. In the Ramen 1 image, four models concentrate on the noodles, while one model focuses on surrounding objects, such as gyoza or a wooden plate. Similarly, in the Ramen 3 image, which features blue-colored soup, four models indicate that it is unattractive. From Pizza 1, four models focus on the toppings, while one model highlights the pizza crust. For Burger 1, four models focus on the fries, while one model focuses on the burger itself. Finally, in Cake 2, two models focus on the strawberry topping, while only one model concentrates on the body of the cake.

To summarize, by applying Grad-CAM, the proposed framework can not only analyze the level of food attractiveness but also identify the locations in the image that make it more or less attractive. This framework reveals insights into food image attractiveness that allow users to choose the most appropriate food images for advertising.

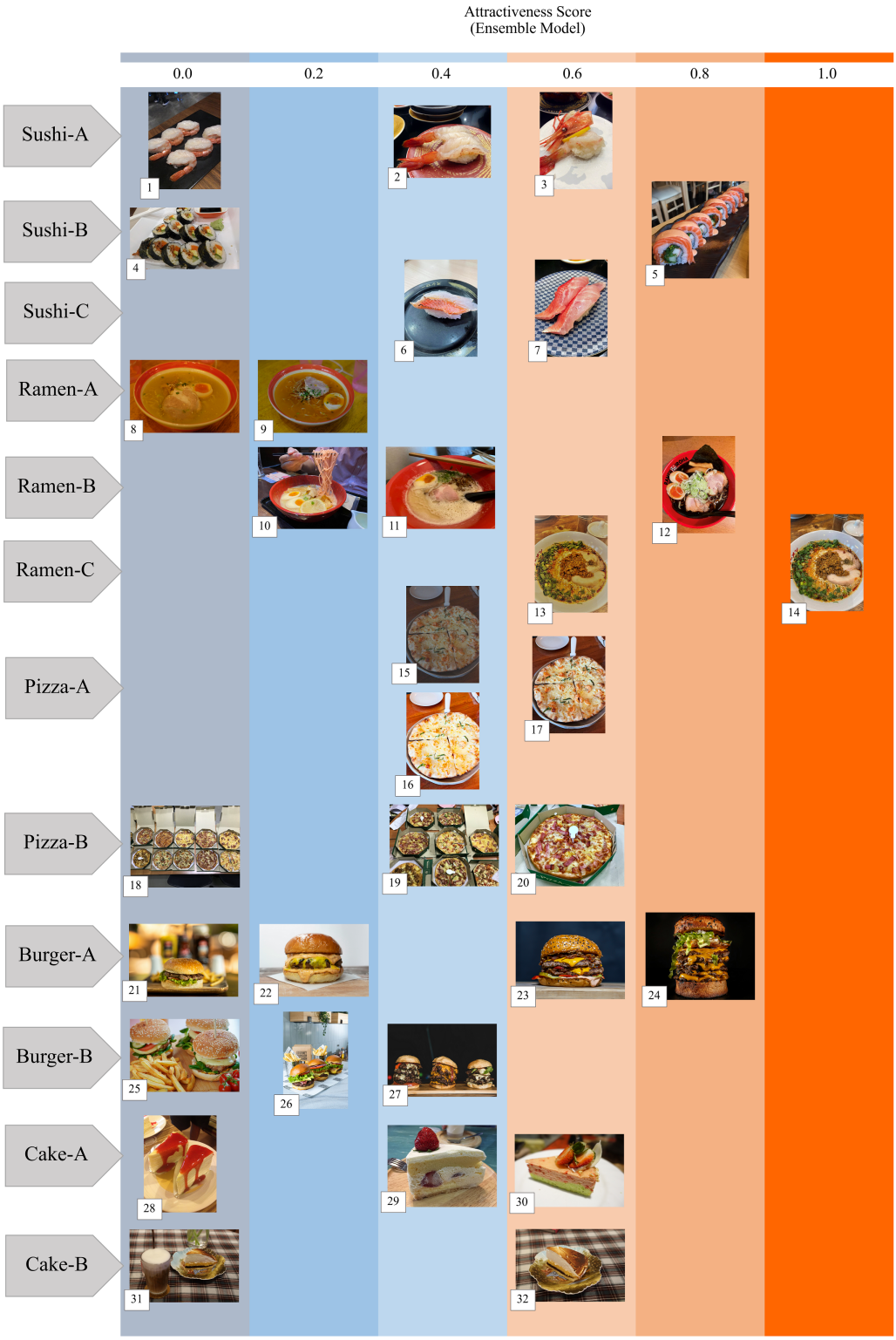

Figure 9. Groups of sample images organized by their categories and close similarity, along with their attractiveness scores. The numbers in the top line indicate the attractiveness scores obtained from our ensemble model. The numbers displayed in the bottom-left corner of each sample image indicate the image index.

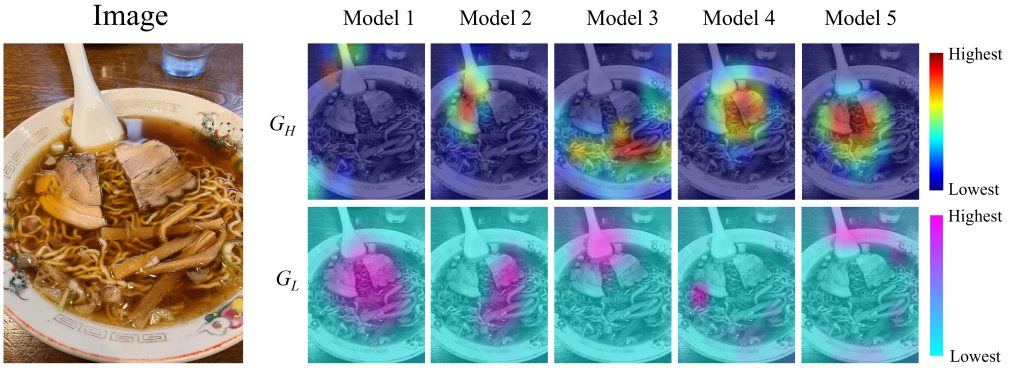

Figure 10. Example of an image analyzed using Grad-CAM from five models.

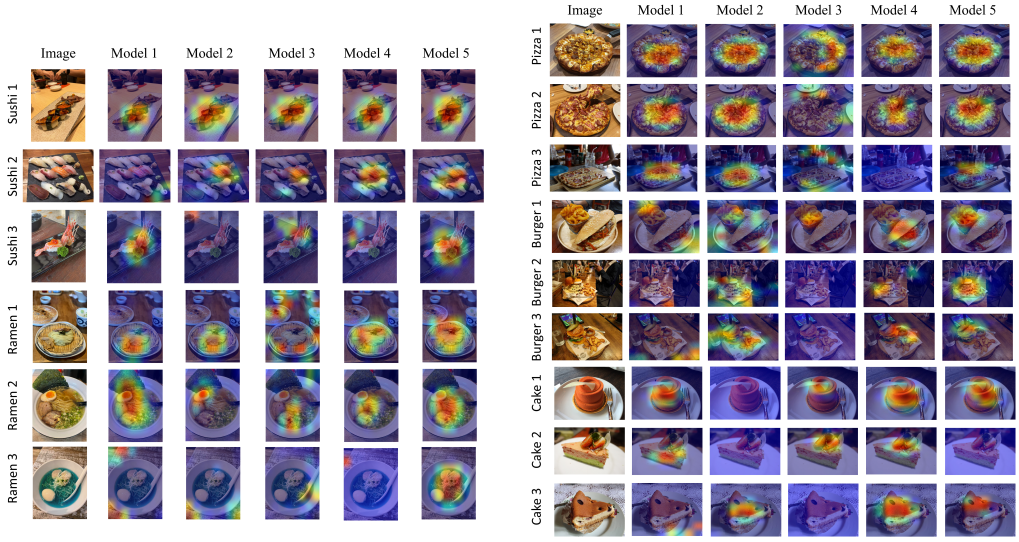

Figure 11. Example of G_H (class H attractiveness) map images from all five categories, analyzed by Grad-CAM across five models.

5. Conclusions

In this paper, we have presented a novel end-to-end framework for assessing the attractiveness of food images without the need for explicit human annotation. By leveraging user engagement data readily available on social media platforms, we have developed a procedure to collect, filter, and partition data in order to mitigate biases stemming from poster identity and posting duration. Highly attractive food images and less attractive food images are selected based on their user engagement level for inclusion in the dataset. We conducted A/B testing with 51 human participants on a total of 250 image pairs from five popular food categories (sushi, ramen, pizza, burger, and cake). The aim was to evaluate the consistency of our proposed engagement-based attractiveness class labeling with the majority of human judgments. The results show that the proposed dataset creation method achieves 97.2% consistency across five food categories. As the final step, we created ensemble models to evaluate the attractiveness of food images using the dataset obtained in the previous step. The ensemble models achieved a consistency of 76.0% (averaging from five food categories) compared to human judgments obtained through A/B testing. In the fine-grained evaluation, the ensemble models' judgments were closer to human judgments than those of the best individual models. Moreover, analyzing the results with Grad-CAM provided valuable insights into the regions that influence attractiveness judgments. These results highlight the potential of our method to capture and encode visual appeal accurately, aligning with diverse public preferences.

The proposed approach offers a scalable and efficient solution for the computational aesthetic assessment of food imagery, circumventing the labor-intensive task of manual annotation. Our method can be applied to various domains and industries where visual appeal plays a crucial role, such as marketing, advertising, and user experience design. Furthermore, our work contributes to a better understanding of visual appeal in the food and hospitality industries, providing valuable insights for businesses and individuals seeking to enhance user engagement and drive consumer behavior through visually appealing food content.

For future work, it is also essential to consider other potential biases not addressed in this study. While our research has taken steps to mitigate certain biases, such as adjusting for basic differences in user engagement, further exploration into factors like paid sponsorships is necessary. These could affect image visibility and, consequently, the attractiveness assessments. Future research should aim to classify engagement among sponsored images separately to maintain the integrity of the attractiveness evaluations. While our experiments focused on specific food categories, the methodology can be extended to other domains and applications where user engagement data are available. Future research could explore the integration of additional features, such as image metadata or textual descriptions, to further enhance the attractiveness assessment models. Additionally, future research could focus on training CNNs to remove duplicated and irrelevant images. This enhancement would enable the framework to be fully automated from the initial photo scraping to the final attractiveness prediction.

Author Contributions: Conceptualization, T.M.; methodology, T.M.; software, T.M.; validation, T.M., K.P. and Y.S.; investigation, T.M.; resources, T.M.; data curation, T.M.; writing—original draft preparation, T.M.; writing—review and editing, K.P. and Y.S.; visualization, K.P.; supervision, K.P. and Y.S.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by Chiang Mai University.

Data Availability Statement: The original contributions presented in the study are included in this article; further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lepkowska-White, E. Exploring the Challenges of Incorporating Social Media Marketing Strategies in the Restaurant Business. *J. Internet Commer.* **2017**, *16*, 323–342. [CrossRef]
2. Needles, A.M.; Thompson, G.M. Social Media Use in the Restaurant Industry: A Work in Progress. *Cornell Hosp. Rep.* **2013**, *13*, 6–17.
3. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A unified embedding for face recognition and clustering. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015. [CrossRef]
4. Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4685–4694. [CrossRef]
5. Uysal, F.; Hardalaç, F.; Peker, O.; Tolunay, T.; Tokgöz, N. Classification of Shoulder X-ray Images with Deep Learning Ensemble Models. *Appl. Sci.* **2021**, *11*, 2723. [CrossRef]
6. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.; van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med Image Anal.* **2017**, *42*, 60–88. [CrossRef] [PubMed]
7. Teeyapan, K. Deep learning-based approach for corneal ulcer screening. In Proceedings of the 12th International Conference on Computational Systems-Biology and Bioinformatics, New York, NY, USA, 14–15 October 2021; pp. 27–36. [CrossRef]
8. Prasad, S.; Singh, P. Medicinal plant leaf information extraction using deep features. *TENCON IEEE Reg. Conf.* **2017**, *11*, 2722–2726. [CrossRef]
9. Olsen, A.; Konovalov, D.A.; Philippa, B.; Ridd, P.; Wood, J.C.; Johns, J.; Banks, W.; Girgenti, B.; Kenny, O.; Whinney, J.; et al. DeepWeeds: A Multiclass Weed Species Image Dataset for Deep Learning. *Sci. Rep.* **2019**, *9*, 2058. [CrossRef] [PubMed]
10. Chang, S.J.; Huang, C.Y. Deep Learning Model for the Inspection of Coffee Bean Defects. *Appl. Sci.* **2021**, *11*, 8226. [CrossRef]
11. Şengür, A.; Akbulut, Y.; Budak, U. Food Image Classification with Deep Features. In Proceedings of the 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, Turkey, 21–22 September 2019; pp. 1–6. [CrossRef]
12. Kagaya, H.; Aizawa, K.; Ogawa, M. Food Detection and Recognition Using Convolutional Neural Network. *ACM Int. Conf. on Multimedia* **2014**, *11*, 1085–1088. [CrossRef]

13. Zhang, Y.; Deng, L.; Zhu, H.; Wang, W.; Ren, Z.; Zhou, Q.; Lu, S.; Sun, S.; Zhu, Z.; Gorriz, J.; et al. Deep Learning in Food Category Recognition. *Inf. Fusion* **2023**, *98*, 101859. [CrossRef]
14. Ruenin, P.; Bootkrajang, J.; Chawachat, J. A System to Estimate the Amount and Calories of Food that Elderly People in the Hospital Consume. In Proceedings of the 11th International Conference on Advances in Information Technology, Bangkok, Thailand, 1–3 July 2020.
15. Agarwal, R.; Choudhury, T.; Ahuja, N.J.; Sarkar, T. Hybrid Deep Learning Algorithm-Based Food Recognition and Calorie Estimation. *J. Food Process. Preserv.* **2023**, *2023*, 6612302. [CrossRef]
16. Takahashi, K.; Doman, K.; Kawanishi, Y.; Hirayama, T.; Ide, I.; Deguchi, D.; Murase, H. Estimation of the attractiveness of food photography focusing on main ingredients. In Proceedings of the 9th Workshop on Multimedia for Cooking and Eating Activities in Conjunction with The 2017 International Joint Conference on Artificial Intelligence, Melbourne, Australia, 20 August 2017; pp. 1–6.
17. Takahashi, K.; Hattori, T.; Doman, K.; Kawanishi, Y.; Hirayama, T.; Ide, I.; Deguchi, D.; Murase, H. Estimation of the attractiveness of food photography based on image features. *IEICE Trans. Inf. Syst.* **2019**, *102*, 1590–1593. [CrossRef]
18. Min, W.; Liu, L.; Wang, Z.; Luo, Z.; Wei, X.; Wei, X.; Jiang, S. Isia food-500: A dataset for large-scale food recognition via stacked global-local attention network. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 393–401.
19. Min, W.; Wang, Z.; Liu, Y.; Luo, M.; Kang, L.; Wei, X.; Wei, X.; Jiang, S. Large scale visual food recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 9932–9949. [CrossRef] [PubMed]
20. Waltner, G.; Schwarz, M.; Ladstätter, S.; Weber, A.; Luley, P.; Lindschinger, M.; Schmid, I.; Scheitz, W.; Bischof, H.; Paletta, L. Personalized dietary self-management using mobile vision-based assistance. In Proceedings of the New Trends in Image Analysis and Processing–ICIAP 2017: ICIAP International Workshops, WBICV, SSPandBE, 3AS, RGBD, NIVAR, IWBAAS, and MADiMa 2017, Catania, Italy, 11–15 September 2017; Revised Selected Papers 19; Springer: Berlin/Heidelberg, Germany; 2017, pp. 385–393.
21. Thung, K.H.; Raveendran, P. A survey of image quality measures. In Proceedings of the 2009 International Conference for Technical Postgraduates (TECHPOS), Kuala Lumpur, Malaysia, 14–15 December 2009; pp. 1–4. [CrossRef]
22. Prasad, S.; Singh, P.P. A compact mobile image quality assessment using a simple frequency signature. In Proceedings of the 2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV), IEEE, Singapore, 18–21 November 2018; pp. 1692–1697.
23. Yang, J.; Lyu, M.; Qi, Z.; Shi, Y. Deep Learning Based Image Quality Assessment: A Survey. *Procedia Comput. Sci.* **2023**, *221*, 1000–1005. [CrossRef]
24. Datta, R.; Joshi, D.; Li, J.; Wang, J.Z. Studying Aesthetics in Photographic Images Using a Computational Approach. In *Proceedings of the Computer Vision—ECCV 2006*; Leonardis, A., Bischof, H., Pinz, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2006; pp. 288–301.
25. Zhang, L.; Gao, Y.; Zimmermann, R.; Tian, Q.; Li, X. Fusion of Multichannel Local and Global Structural Cues for Photo Aesthetics Evaluation. *IEEE Trans. Image Process.* **2014**, *23*, 1419–1429. [CrossRef] [PubMed]
26. Lu, X.; Lin, Z.; Jin, H.; Yang, J.; Wang, J.Z. Rating Image Aesthetics Using Deep Learning. *IEEE Trans. Multimed.* **2015**, *17*, 2021–2034. [CrossRef]
27. Deng, Y.; Loy, C.C.; Tang, X. Image Aesthetic Assessment: An experimental survey. *IEEE Signal Process. Mag.* **2017**, *34*, 80–106. [CrossRef]
28. Yang, H.; Shi, P.; He, S.; Pan, D.; Ying, Z.; Lei, L. A Comprehensive Survey on Image Aesthetic Quality Assessment. In Proceedings of the 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS), Beijing, China, 17–19 June 2019; pp. 294–299. [CrossRef]
29. Pu, Y.; Liu, D.; Chen, S.; Zhong, Y. Research Progress on the Aesthetic Quality Assessment of Complex Layout Images Based on Deep Learning. *Appl. Sci.* **2023**, *13*, 9763. [CrossRef]
30. Sheng, K.; Dong, W.; Huang, H.; Ma, C.; Hu, B.G. Gourmet photography dataset for aesthetic assessment of food images. In Proceedings of the SIGGRAPH Asia 2018 Technical Briefs, Tokyo, Japan, 4–7 December 2018; pp. 1–4.
31. Sheng, K.; Dong, W.; Huang, H.; Chai, M.; Zhang, Y.; Ma, C.; Hu, B.G. Learning to assess visual aesthetics of food images. *Comput. Vis. Media* **2021**, *7*, 139–152. [CrossRef]
32. Philp, M.; Jacobson, J.; Pancer, E. Predicting social media engagement with computer vision: An examination of food marketing on Instagram. *J. Bus. Res.* **2022**, *149*, 736–747. [CrossRef]
33. Attokaren, D.J.; Fernandes, I.G.; Sriram, A.; Murthy, Y.S.; Koolagudi, S.G. Food classification from images using convolutional neural networks. In Proceedings of the TENCON 2017 IEEE Region 10 Conference, IEEE, Penang, Malaysia, 5–8 November 2017; pp. 2801–2806.
34. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826. [CrossRef]
35. Islam, M.T.; Siddique, B.N.K.; Rahman, S.; Jabid, T. Food image classification with convolutional neural network. In Proceedings of the 2018 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), IEEE, Bangkok, Thailand, 21–24 October 2018; Volume 3, pp. 257–262.

36. Singla, A.; Yuan, L.; Ebrahimi, T. Food/Non-food Image Classification and Food Categorization using Pre-Trained GoogLeNet Model. In Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management, Amsterdam, The Netherlands, 16 October 2016.
37. Shimoda, W.; Yanai, K. Learning Food Image Similarity for Food Image Retrieval. In Proceedings of the 2017 IEEE Third International Conference on Multimedia Big Data (BigMM), Laguna Hills, CA, USA, 19–21 April 2017; pp. 165–168. [CrossRef]
38. A, K.; Lanke, R. Image Retrieval based on Deep Learning–Convolutional Neural Networks. In Proceedings of the 2022 International Interdisciplinary Humanitarian Conference for Sustainability (IIHC), Bengaluru, India, 18–19 November 2022, pp. 757–762. [CrossRef]
39. Singh, P.P.; Prasad, S. A Hybrid Adaptive Image Retrieval Approach by Using Clustering and Neural Network Techniques. In *Intelligent Data Engineering and Analytics*; Bhateja, V. Yang, X.S. Chun-Wei, L.J., Das, R., Eds.; Springer: Singapore, 2023; pp. 395–403.
40. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
41. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
42. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
43. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
44. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
45. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626. [CrossRef]