

# Integrating Computational Metrics and Human Judgments with Deep Learning to Enhance Website Aesthetic Evaluation

Nehal Ali Ataif<sup>1</sup>, Abdulaziz Attaallah<sup>2</sup>, Rsha Mirza<sup>3</sup>, Emad Sami Jaha<sup>4</sup>

Submitted:13/03/2024    Revised: 28/04/2024    Accepted: 05/05/2024

**Abstract:** The visual aesthetics of web pages are the most important aspects that capture users' attention at first glance. The design's aesthetic plays an important role, for example, when subscribing to the site or purchasing online services. However, developing a design that satisfies all users is a major challenge due to users' different tastes, cultures, and experiences. In this context, computational methods have proven efficient in assessing measurable objective aesthetics but are unable to assess the personal aesthetics the human eye captures. Human judgment can evaluate subjective aesthetics, but it is inconsistent. Deep learning-based solutions often overlook multiple perspectives, focusing mainly on subjective satisfaction. In this paper, we present a deep learning based integrated approach that integrates computational methods and human evaluation to evaluate different aspects of visual aesthetics of web pages. Six objective aesthetic attributes were measured using the computational method to train a baseline model based on multiple regression. The base model was fine-tuned to train a model on three subjective aesthetics evaluated by human designers. Our approach achieved a mean squared error of (0.015, 0.023, 0.035) and excellent correlation with human evaluation (0.931, 0.859, 0.841) for the subjective ratings (rate consistency, rate clarity, and satisfaction). These results demonstrate the effectiveness of our proposed approach in simultaneously automating the visual design of various aspects of webpage aesthetics through multi-output regression.

**Keywords:** Aesthetics, Webpage, Computational Methods, Deep Learning, Subjective, Objective, Multi-output Regression

## 1. Introduction

In the digital space today, the look of a website could make all the difference. It is not about being pretty; it can affect how the user feels and acts. With the continuous crowding over the internet, outstanding, unique, and attractive design has become necessary [1]. The aesthetics of websites are the aspect that captivates users' attention, creates a good impression on the very first look, and helps retain them for a long time. A website that's easy on the eyes can draw more visitors, keeping them longer and boosting those conversion rates [2]. However, judging a website's aesthetic value is a challenging task. The challenge remains: how to balance these complex scores against human opinions. Traditional methods rely much on automated metrics, which are excellent for assessing measurable aesthetics but often lack the subtleties that human eyes can capture [3]. Metrics are computational models gauging aesthetics by assessing symmetry, balance, complexity, contrast, and harmony. Admittedly, they are quick to make judgments and be very consistent. However, they usually fail to capture subjective and contextual elements that shape human perception and preference [4]. On the other hand, human assessments lack the consistency and stability of automated tools, even though they are full of subjective insights [2]. Human

evaluation clearly records individual and cultural preferences based on user feedback such as ratings, rankings, comparisons, or comments [5], [6]. However, these approaches typically have high variability and low reliability, making it difficult and costly to achieve the level of reliability required to accurately assess the overall aesthetics of a website [7]. Recent deep learning development, particularly with CNNs, has effectively estimated the aesthetic appeal of images [8], [9], [10], [11]. This superior can replace manually designed features with fully automatic learning systems to achieve more efficient solutions in aesthetics-related tasks [9]. While webpages are significantly different from traditional images, many studies have proved that CNN can go beyond the performance of previous methods based on hand-crafted features [12], [13], [14]. However, these methods often overlook multiple perspectives, focusing mainly on subjective satisfaction. Moreover, they leverage transfer learning on pre-trained models from different domain datasets, such as ImageNet de et al. [15] and Flickr style Delitzas et al.[12] posing challenges for effectively applying transfer learning in website aesthetics. Notably, there are pre-trained models trained on small datasets based on user ratings [12], which may be affected by high variability and low reliability. Conversely, computational metrics like the QUESTIM tool [16] provide quick and consistent scores for measurable aesthetics. Thus, the integration of computational metrics with human judgments remains an underexplored area in the existing literature. In this paper, we introduce a novel deep learning approach that combines computational metrics

<sup>1234</sup> Department of CS, Faculty of Computing and Information Technology, KAU, Jeddah, KSA

ORCID ID : 0009-0008-9556-4074

ORCID ID : 0000-0001-5363-040X

ORCID ID : 0000-0003-4217-647X

ORCID ID : 0000-0002-3184-5466

\* Corresponding Author Email: nahmedataif@stu.kau.edu.sa

with human evaluations to enhance web page aesthetics evaluation. Initially, we use the QUESTIM tool to score various aspects of website aesthetics (six objective aesthetics) on a diverse dataset of 1000 website screenshots covering a wide range of web categories. This dataset is used as a source domain to train an EfficientNetB0-based multi-output regression model as the base model. Subsequently, we fine-tune the base model by training it on a new dataset (target domain) of website screenshots evaluated by human designer experts, focusing on three subjective aesthetics. The following are the paper's primary contributions:

- A unique dataset of 1,000 website screenshots annotated on six distinct objective metrics using the QUESTIM tool and three subjective aesthetics by human experts.
- We introduced a novel deep learning-based approach that integrates computational metrics and human evaluations to automate the assessment of various aspects of webpage aesthetics through multi-output regression.

Our integrated approach to website aesthetic evaluation has proven effective, as evidenced by the results. Substantial performance improvements were observed when we trained a base model on objective aesthetics metrics obtained from the QUESTIM tool and then refined it on a dataset assessed by human experts. It was shown that integrating objective metrics aided in the learning of more resilient aesthetic representations, as the optimized model routinely outperformed the one trained from scratch. Through further refinement through expert human judgments, this integrated approach not only improves the model's accuracy but also leads to a more thorough evaluation of website aesthetics. The remainder of the paper is organized as follows: Section 2 provides a critical summary of previous studies and points out gaps in the body of knowledge. Section 3 explains the proposed methodology in detail, including dataset acquisition, annotation methods, deep learning models and their configurations, and evaluation metrics. In Section 4, the experimental results are presented along with a thorough analysis of the model performance, including performance metrics and visualizations. Result interpretation, discussion and a comparison with the existing approach in the literature are given in Section 5. The paper concludes in Section 6 with recommendations for future work.

## 2. Related Work

Over the past decade, evaluating user interface (UI) aesthetics has received significant attention. Various approaches have been proposed, which can be broadly divided into three categories: computation-based methods, user perception-based methods, and deep learning-based methods. In this section, we will explore these categories and highlight the advantages and disadvantages of each approach.

### 2.1. Computation metric-Based Website Aesthetics Evaluation

Computational metric methods assess the aesthetic impact on user interfaces (UIs) using mathematical or geometric principles. They are based on the balance of symmetry, proportion consistency, simplicity, and regularity and are developed from visual design methods. The purpose of geometric-based evaluation is to bring an objective and consistent method to evaluate UI aesthetics. This could significantly impact users' experience, usability, trust, and satisfaction. Zen et al. introduced the QUESTIM (Quality Estimator using Metrics) tool, which evaluates graphical user interfaces using quantifiable metrics [16]. QUESTIM provides a definition of a manual area of interest definition on a webpage or screenshot and then allows calculations based on these selections. This tool provides a metrics report that sheds light on various aesthetic components including, but not limited to, balance, sequence, equilibrium, unity, symmetry, proportion, economy, density, homogeneity, regularity, rhythm, and order. The pilot study conducted with such metrics by the authors had significant effects on users' perceptions of UI design properties; therefore, metrics-based evaluation has the potential to provide objective feedback to designers for making an aesthetically pleasing interface while keeping its original design concept intact. However, despite their advantages, geometric-based methods face a variety of problems. By their very nature, they might not truly capture the semantic and emotional aspects of UI aesthetics. They can also be susceptible to resolutions and size of UI designs [17]. Furthermore, the subjectivity of human perception could influence the evaluation due to taste, cultural background, and contextual factors that shape aesthetic judgment [17].

### 2.2. User Perception-Based Website Aesthetics Evaluation

User perception-based approaches focus on understanding how users rate the attractiveness of a website and interpreting these opinions to influence user behavior and satisfaction. This method collects user opinions and preferences on various aesthetic elements such as simplicity, fonts, layout, color, and craftsmanship by asking them to rate the aesthetics of the website. Thielsch et al.[2] examined user perception-based evaluations by examining how the user's visual perception affects their interaction with websites. The study showed the significance of first impressions and how the initial visual appeal of a website can impact long-term user engagement. Miniukovich et al. examined some users' cognitive and emotional responses to website designs and found that users' aesthetic perceptions are influenced by individuals' preferences, cultural backgrounds, emotional states, and cognitive processes [18]. However, these methods also have their disadvantages. Most of these methods are time-consuming, expensive, and

prone to variations and biases [19]. Different studies use different methods, measures, and models, which may lead to inconsistencies in the results of website aesthetics assessment. For example, some studies use subjective user ratings based on questionnaires or scales [20], while others use objective data such as eye tracking or physiological measurements [21]. Furthermore, ratings based on users' perceptions would vary over time due to the duration and frequency of exposure to a website [6]. Despite user perception-based methods providing a crucial perspective in evaluating website aesthetics and capturing diverse and subjective experiences that objective metrics may miss, these approaches are fraught with challenges such as high cost, time commitment, and susceptibility to bias and inconsistency.

### 2.3. Deep Learning-Based Aesthetics Evaluation

Different deep learning-based methods have been introduced to predict GUI visual aesthetics as a result of machine learning advancements [9], [12], [13], [15], [22]. To cut down on training time, some methods use pre-trained networks. Khani et al. In [14], they constructed their model using the AlexNet architecture and trained it with 418 website screenshots that were rated according to their visual appeal. Their approach involved a hybrid model that integrated traditional machine learning methods with convolutional neural networks (CNN). The model classifies each image's visual aesthetics as "good" or "bad" using a Gaussian radial basis function, which is generated by a support vector machine. They report an error rate (root mean squared error) of 34.15 percent as a result of the approach. Dou et al utilized 398 website snapshots for training their model, employing a pre-trained CNN (CaffeNet) [13]. Unlike Khani et al., their dataset was categorized not by two classes but by the average score of human ratings on a 9-point scale, where 1 indicated the lowest visual appeal, and 9 represented the highest. They approached predicting visual aesthetics as a regression task, with labels ranging from 1 to 9. Their reported error rate was 20.41%. A different method was proposed by Xing et al. [23], who trained five models using 38,423 GUI images sourced from a popular Chinese website for UI designers. These models were labeled based on the number of 'likes' and user 'collections' associated with the user interfaces. They discovered that a Squeeze and-Excitation VGG model performed the best, achieving error rates of 89% for 'likes' and 38% for 'collections'. However, they did not propose a strategy to merge these two distinct outcomes into a single, standalone visual aesthetics value. Despite their potential, current approaches to deep learning-based aesthetics assessment face several challenges. These methods often rely on human-annotated datasets and have inconsistencies and high variability, compromising the quality of data that is essential for developing robust models. Additionally, while some strategies use transfer learning with ImageNet

weights or tailor models specifically for user interface aesthetics, the difference in domains can impact their effectiveness. To address these issues, we propose a novel framework that uses a dataset annotated with automated metrics across six aesthetic dimensions using the QUETIM tool, known for its consistency and reliability. This foundation enables the initial training of our base deep learning model. We further refine this model by refining its deeper layers with human-rated aesthetics and integrating the precision of automated metrics with the depth of human judgment. Such an approach not only improves the quality and diversity of datasets, but also ensures that training is closely aligned to specific aesthetic assessments, thus bridging the gap between automated tools and human insights. This methodology promises a more comprehensive and nuanced understanding of website aesthetics tailored to a wide range of user preferences.

## 3. The Proposed Approach

Aiming to enhance the assessment of visual aesthetics on websites, we developed a novel approach that combines automated metrics with human evaluations. This approach uses a transfer learning model based on multi-output regression. The approach we use consists of several sequential steps. First, we created a dataset by collecting scrolling screenshots from numerous categories of websites. We then labeled nine aesthetic features, six of which were evaluated using the QUESTIM tool and the remaining three by human designers. The integration of these nine aesthetic elements into the final dataset was made possible by the thorough annotation process. We used a modified version of the EfficientNet B0 architecture to apply the integrated data. EfficientNet is a Convolutional Neural Network (CNN) that achieves scalability and efficiency by using a composite scaling technique [24], [25]. This technique adjusts the width, depth, and resolution of the network evenly, taking into account the available resources, which allows the processing of visual appeals with a balance between accuracy and efficiency. The base model was trained using the six aesthetics evaluated with QUESTIM. The model was then refined based on the three aesthetics evaluated by humans. Fig 2 shows the proposed model.

### 3.1. Dataset Collection and Preparation

To collect the dataset, we captured scrolling screenshots from a wide range of website categories in portrait mode. All screenshots were pre-processed to standardize the screenshot sizes.

#### 3.1.1. Data Collection

We collected 1000 portrait-oriented scrolling screenshots. We took care to gather a varied dataset that encompassed a variety of website categories, such as corporate, government, e-commerce, membership, and online forums. We collected screenshots and saved it as 922 x 744-pixel

JPG images using the GoFullPage tool [30]. The dataset collection process resulted in 1678 images gathered from 963 websites. We applied particular criteria to eliminate the screenshots in order to guarantee unique and varied aesthetics while also ensuring the dataset's relevance and focus. Table 1 explains the website's screenshots collected and selected. 1000 screenshots were chosen to ensure relevance and focus, with 200 images curated for each category.

**Table 1.** Summary of screenshot collection and selection by website category

Category	Collected	Selected
<b>E-commerce</b>	383	200
<b>Membership</b>	347	200
<b>Online Forums</b>	264	200
<b>Business</b>	427	200
<b>Governmental</b>	257	200
<b>Total</b>	1678	1000

### 3.1.2. Dataset Annotation

This section details the annotation process for preparing the collected dataset for training our deep learning models. The QUESTIM tool was used to measure objective aesthetics, while human designer experts evaluated subjective assessments.

**QUESTIM based annotation:** For the objective evaluation, we utilized the QUESTIM tool [16] to assess six aesthetic metrics. These metrics, chosen for their relevance in capturing various aspects of visual aesthetics of a screenshot image ( $I$ ), were evaluated on a scale from 0 to 1. Specifically, when interpreting data metrics where 0 indicates low and 1 indicates high, it is important to note that this is standard for all metrics except compression complexity. For compression complexity, the scale is inverted. The aesthetics metrics evaluated by QUESTIM are:

1. Saliency Balance (**SB**): Measures the balance of visual saliency across the webpage.
2. Border Balance (**BB**): Assesses the distribution of borders within the layout.
3. Border Density (**BD**): Quantifies the density of borders in the design.
4. Color Density (**CD**): Evaluates the distribution and concentration of colors.

5. Colorfulness (**CF**): Measures the overall colorfulness of the design.
6. Compression Complexity (**CC**): Assesses the complexity of the visual design in terms of compression.

These metrics can be formally represented as a vector. For each website screenshot  $i$  we have:

$$Q_i = [SB, BB, BD, CD, CF, CC] \in [0,1]^6$$

**Human expert-based annotation:** In addition to automated evaluation, we incorporated human judgments into our assessment process. A panel of human experts evaluated each design based on three additional aesthetic metrics, rated on a scale from 1 to 5. The aesthetics metrics evaluated by human experts are:

1. Rate Consistency (**RCS**): Evaluates the consistency of the visual design.
2. Rate Clarity (**RC**): Assesses the clarity and readability within website design.
3. Rate Satisfaction (**RS**): Measures the overall satisfaction with the design.




These human evaluations can be represented as another vector for each website screenshot  $i$ :

$$H_i = [RCS, RC, RS] \in [1,5]^3$$

Combining both the automated metrics and human evaluations, each website screenshot  $i$  is represented by a vector that integrates both objective and subjective scores:

$$A_i = [SB, BB, BD, CD, CF, CC, RCS, RC, RS]$$

where  $A_i \in \mathbb{R}^9$ .

		
SB: 0.88	SB: 0.81	SB: 0.62
BB: 0.93	BB: 0.73	BB: 0.73
BD: 0.24	BD: 0.18	BD: 0.07
CD: 0.63	CD: 0.54	CD: 0.05
CF: 0.98	CF: 0.57	CF: 0.06
CC: 0.44	CC: 0.49	CC: 0.34

RCS: 3	RCS: 4	RCS: 2
RCT: 4	RCT: 5	RCT: 3
RS: 5	RS: 3	RS: 4

**Fig 1.** Examples of annotation processes using QUESTIM and Human experts.

The nine aesthetic metrics were recorded in an Excel sheet as shown in Fig 1, where each line referred to a single website screenshot  $I$ . In such a way, the structure of this dataset provided the base for a widely applicable evaluation framework; it enabled the combination of objective metrics stemming from the QUESTIM tool with subjective insights obtained from human experts.

### 3.1.3. Data Preprocessing

All images were resized to a resolution of  $512 \times 384$  pixels to standardize input dimensions across all images. Additionally, pixel intensity values were normalized to fall within the range of 0 to 255, facilitating efficient training and improving model convergence [26]. The dataset was split into training and validation sets using an 80:20 ratio, resulting in 800 images for training and 200 images for validation. In addition, 100 new screenshots were collected and annotated in the same way. The test set is used to evaluate the model.

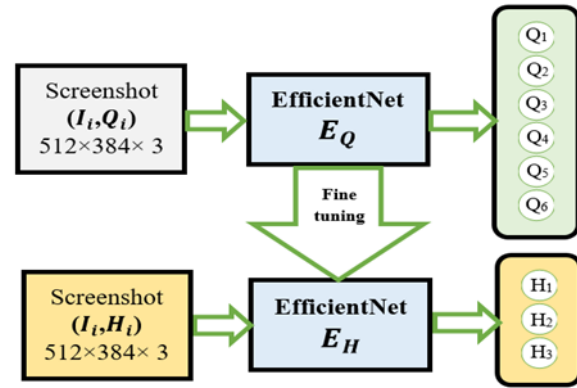
### 3.2. EfficientNet-based Multi-Output Regression

In the domain of UI aesthetics, the application of deep learning models, particularly convolutional neural networks (CNNs), has gained significant attention for their ability to extract and learn complex features from visual inputs effectively [8], [9], [15], [22]. In this section, we explain, the structure and hyperparameters of models used in this paper.

#### 3.2.1. Model Structure

Convolutional Neural Networks (CNNs) have developed significantly in terms of both accuracy and computational requirements. EfficientNet is one of the greatest developments that effectively meets the increasing demand for high performance and computing requirements, introduced by Tan and Le in 2019 [24], optimizes the network architecture in depth, width and resolution using a composite scaling technique. This method is particularly useful in situations where resources are limited, as it not only improves the accuracy of the model but also dramatically reduces the computing power required. In this paper, we leveraged the EfficientNet B0 architecture, initially designed for image classification [24] since it is shown from previous works to work quite suitably concerning Mobile UI aesthetics assessment tasks [15], [27]. The final layer was updated with a six-neuron, three-

neuron, and nine-neuron for objective subjective and integrated aesthetics, respectively, where each neuron corresponds to one of the aesthetic metrics, following the same principle as [12], [15]. This enabled the model to learn several aesthetic scores simultaneously. We explain our approach as shown in Fig 2.



**Fig 2.** Architecture of the Proposed Approach: The diagram illustrates our dual-phase model training process using the EfficientNet architecture.  $E_Q$  represents the base model configured with an output layer of six neurons, each corresponding to one of the six objective aesthetic metrics ( $Q_1, Q_2, Q_3, Q_4, Q_5, Q_6$ ). This base model  $E_Q$  was trained on the pairs  $(I_i, Q_i)$  to capture features associated with objective aesthetics. Subsequently  $E_H$ , which inherits weights from  $E_Q$ , undergoes fine-tuning on its initial seven layers to adapt to features related to the three subjective aesthetics ( $H_1, H_2, H_3$ ).

Let's denote the base model as  $E_Q$ , which is built upon the EfficientNet architecture. The output layer of  $E_Q$  is adapted to have six neurons, corresponding to six objective aesthetics  $\{Q_1, Q_2, Q_3, Q_4, Q_5, Q_6\}$ . The model  $E_Q$  is defined by its parameters  $\theta_Q$ . We first started to learn the base model on objective aesthetics. The input to  $E_Q$  consists of image pairs  $(I_i)$  with corresponding objective aesthetics labels  $Q_i = (q_{i1}, q_{i2}, q_{i3}, q_{i4}, q_{i5}, q_{i6})$ . The loss function  $L_Q$  used for training  $E_Q$  is typically a mean squared error (MSE) calculated as:

$$L_Q(\theta_Q) = \frac{1}{n} \sum_{i=1}^n \| E_Q(I_i; \theta_Q) - Q_i \|^2 \quad (1)$$

Where  $n$  is the number of training samples. After that, we applied transfer learning for subjective aesthetics. Let  $E_H$  denotes the model after transferring weights from  $E_Q$  and modifying it for subjective aesthetics with three additional outputs  $\{H_1, H_2, H_3\}$ . Only the seven earlier layers of  $E_H$  are fine-tuned, with the parameters for these layers denoted as  $\theta_H$ . The adapted model  $E_H$  is trained on a new set of image pairs  $(I_i)$  with corresponding subjective labels  $H_i = (h_{i1}, h_{i2}, h_{i3})$ . The loss function  $L_H$  for  $E_H$  could also be an MSE, defined as:

$$L_H(\theta_H) = \frac{1}{m} \sum_{i=1}^m \|E_H(I_i; \theta_H) - H_i\|^2 \quad (2)$$

Where  $m$  is the number of samples in the subjective training dataset. The mean-squared error (MSE) was used as the loss function during training. Stochastic gradient descent (SGD) optimization was done with an initial learning rate of 0.001. Another technique to further improve the training includes early stopping, learning rate reduction on a plateau, and checkpoint monitoring for preventing overfitting and ensuring optimal model performance.

### 3.2.2. Fine-Tuning Sitting

The fine-tuning process adapts a pre-trained EfficientNetB0 model, initially trained on a dataset scored with objective metrics, to a new dataset evaluated by human experts. This approach leverages the strengths of both computational metrics and human evaluations to create a model that captures a more comprehensive view of website aesthetics. The layers from block6d\_se\_excite onward are set to be trainable. These layers contain high-level features specific to aesthetics' evaluation, allowing the model to better align with human expert judgments. In addition, Batch-Normalization layers are kept frozen to prevent destabilization. Keeping them frozen ensures that the running statistics (mean and variance) learned during initial training remain consistent.

### 3.3. Evaluation Metrics

The main goal of our approach is to explore the effectiveness of an integrated approach. Therefore, we used the Pearson Correlation Coefficient (PCC) to evaluate the performance of our models in terms of the correlation of predicted scores with web designers' evaluation. We calculate the p-values and the 95% confidence intervals as Bonett et al. [28]. A p-value of less than 0.05 is considered statistically significant, as supported by the findings in [29]. This study reports that all PCC values achieve statistical significance, with p-values below 0.001. As PCC is limited to assessing linear relationships, we utilize the Root Mean Square Error RMSE to provide a measure of the magnitude of error between the model predictions and the actual scores. For a dataset with  $n$  samples where  $\hat{Q} = E_Q(I_i; \theta_Q)$  predicts objective aesthetics and  $\hat{H} = E_H(I_i; \theta_H)$  predicts subjective aesthetics and  $Q_i$  and  $H_i$  are the actual objective and subjective aesthetic values, respectively. The evaluation metrics used in this paper are defined in (3) to (6).

$$\begin{aligned} PCC_Q &= \frac{\sum_{i=1}^n (Q_i - \bar{Q}_i)(\hat{Q}_i - \bar{\hat{Q}}_i)}{\sqrt{\sum_{i=1}^n (Q_i - \bar{Q}_i)^2} \sqrt{\sum_{i=1}^n (\hat{Q}_i - \bar{\hat{Q}}_i)^2}} \quad (3) \end{aligned}$$

$$\begin{aligned} PCC_H &= \frac{\sum_{i=1}^n (H_i - \bar{H}_i)(\hat{H}_i - \bar{\hat{H}}_i)}{\sqrt{\sum_{i=1}^n (H_i - \bar{H}_i)^2} \sqrt{\sum_{i=1}^n (\hat{H}_i - \bar{\hat{H}}_i)^2}} \quad (4) \end{aligned}$$

$$\begin{aligned} MSE_Q &= \frac{1}{n} \sum_{i=1}^n (Q_i - \hat{Q})^2 \quad (5) \end{aligned}$$

$$\begin{aligned} MSE_H &= \frac{1}{m} \sum_{i=1}^m (H_i - \hat{H})^2 \quad (6) \end{aligned}$$

$$RMSE_Q = \sqrt{MSE_Q} \text{ and } RMSE_H = \sqrt{MSE_H}$$

## 4. Experimental Results

The primary objective of our study was to develop and evaluate a novel framework that integrates automated metrics and human judgments to enhance the assessment of objective and subjective aesthetics in website design. Our main findings indicate that the base model trained on subjective aesthetics that was scored by QUESTIM forms a good source domain to be fine-tuned on subjective aesthetics evaluated by humans. Our approach achieved a higher correlation across different subjective aesthetics. This demonstrates that our approach provides a more accurate and reliable assessment of website aesthetics compared to models that are trained from scratch or fine-tuned on ImageNet. In this experiment, all models' structures are based on EfficientNet B0 with an updated output layer with six, three, and nine neurons Q, H, and A aesthetics, respectively. In addition, all models were tested on 100 screenshots of website UIs fully annotated by human experts for all nine aesthetic metrics. Moreover, the evaluation metrics MSE, RMSE, and PCC were used to evaluate the performance of each model.

### 4.1. QUESTIM (Q) based Model ( $E_Q$ )

During this phase, we trained the base model on six key aesthetic metrics (Q). Table 2 presents the model's performance across six objective aesthetics. The evaluation metrics (MSE, RMSE, and PCC) provide insights into the correlation of the model's predictions with QUESTIM tool scores, which evaluate various aesthetic aspects of website screenshots.

Table 2. Performance of the base model on Q aesthetics metrics (objective aesthetics)

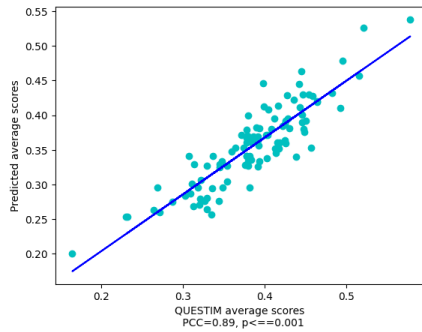
**Table 2.** Performance of the base model on Q aesthetics metrics (objective aesthetics)

Metric	MSE	RMSE	PCC
SB	0.012	0.109	0.636

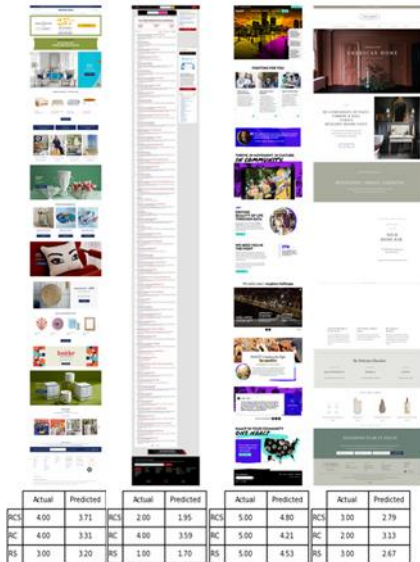


BB	0.008	0.088	0.655
BD	0.006	0.080	0.519
CD	0.005	0.073	0.868
CF	0.007	0.081	0.887
CC	0.002	0.046	0.724

The base model demonstrated promising performance across various aesthetic attributes. Metrics such as MSE, RMSE, and PCC indicated relatively low error rates and strong correlations with Questim scores, particularly noteworthy for attributes like color density and colorfulness. These results underscore the effectiveness of leveraging objective metrics for training deep learning models, providing a robust foundation for subsequent fine-tuning processes.



**Fig 3.** Correlation between QUESTIM scores and  $E_Q$  predictions



**Fig 4.** Performance Evaluation of the  $E_H$  Model on Testing Set: A Multi-Output Regression Analysis: The  $E_H$  model is evaluated on a sample of screenshots from the testing set, assessing three subjective aesthetics: rate consistency (RC), rate clarity (RC), and rate satisfaction (RS). The Fig demonstrates the model's ability to predict these subjective metrics, highlighting its effectiveness in capturing nuanced

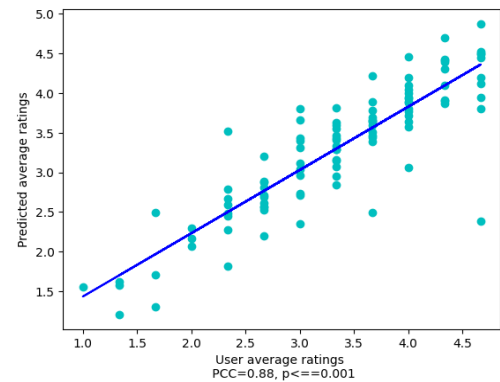
human aesthetic evaluations through fine-tuning.

#### 4.2. Fine-Tune the base model ( $E_Q$ ) on Aesthetics (H)

The base model exhibited promising performance across various categories of aesthetic characteristics. The metrics, including MSE, RMSE, and PCC, demonstrated low error rates and strong positive correlations with Questim scores. This was particularly remarkable for features like color density and colorfulness. The results highlight the effectiveness of utilizing objective aesthetics as a source domain to establish a strong foundation for subsequent fine-tuning on subjective aesthetics evaluated by humans.

**Table 3.** Performance of fine-tuned (base model) on Human based aesthetics evaluation (subjective aesthetics)

Metric	MSE	RMSE	PCC
Rate consistency	0.015	0.124	0.931
Rate clarity	0.023	0.152	0.859
Satisfaction	0.035	0.187	0.841



**Fig 5.** Correlation between the user evaluations and  $E_H$  model predictions (fine-tuning using  $E_Q$ )

As shown in Fig 5, with a PCC of 0.88, the fine-tuned model demonstrates a significant alignment with human judgments.

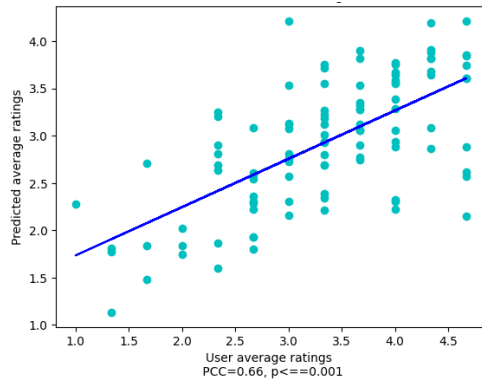
#### 4.3. ImageNet-Based $E_H$ Model

When we fine-tuned the EfficientNet B0 model using ImageNet weights from scratch, we observed lower performance.

**Table 4.** Fine-Tuned from Scratch (ImageNet) on Human based aesthetics evaluation (subjective aesthetics)

Metric	MSE	RMSE	PCC
Rate consistency	0.052	0.229	0.770
Rate clarity	0.061	0.248	0.501
Satisfaction	0.092	0.303	0.549

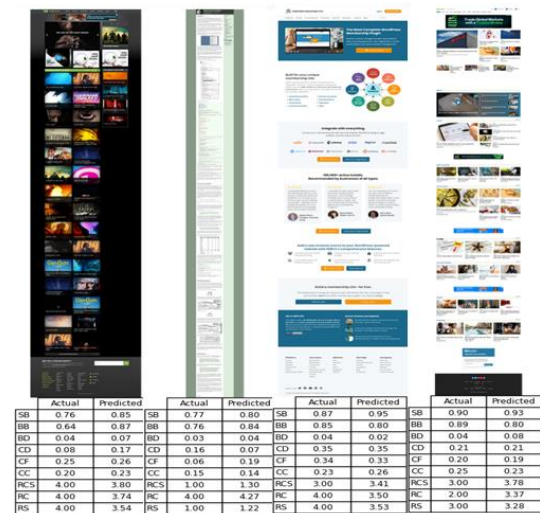
As shown in Table 4, model showed higher error rates, as measured by MSE and RMSE, and lower PCC for all assessed aesthetics compared to the fine-tuned model using EQ. This suggests that while using ImageNet weights can be beneficial for learning general features, it may not capture the subtle details required for web aesthetic assessments as effectively as starting with Questim-based features.



**Fig 6.** Correlation between the user evaluations and  $E_H$  model predictions (fine-tuning using ImageNet from scratch)

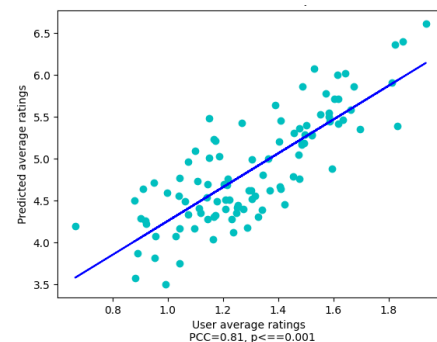
#### 4.4. Fine-tune on integrated model

The fine-tuning of the model on integrated objective (Q) and subjective (H) aesthetics metrics demonstrates notable improvements in the model's ability to evaluate website aesthetics. As shown in Table 5, the model's performance in predicting saliency balance shows a moderate correlation with the actual values, suggesting that although the model can capture some aspects of visual attention distribution, there is still room for improvement. Furthermore, the lower MSE and RMSE values reflect the better performance of the model in predicting border balance. However, the PCC indicates a moderate correlation, suggesting that while the model predicts border balance relatively well, it could benefit from further fine-tuning. Furthermore, the model has good boundary density prediction accuracy with low error values. Nevertheless, the correlation remains moderate, suggesting that the capture of the intricacies of boundary density requires further refinement. In contrast, the high PCC value for color density indicates a strong correlation between the predicted and actual values and shows the robustness of the model in evaluating this metric. Likewise, the model shows strong performance in predicting colorfulness with a high correlation coefficient, suggesting that it accurately assesses the vibrancy and richness of colors in website aesthetics. In addition, the performance of the model in predicting compression complexity is also strong (low) error values and a high PCC.



**Fig 7.** Performance Evaluation of the  $E_{HQ}$  Model on Testing Set: A Multi-Output Regression Analysis: The  $E_{HQ}$  model is evaluated on a sample of screenshots from the testing set, assessing six objective aesthetics and three subjective aesthetics

As illustrated in Fig 8, which visualizes the average correlation with a PCC of 0.81 between user ratings and the model's predictions, the fine-tuned model achieves a significant alignment with human judgments. This high correlation underscores the model's potential to effectively bridge the gap between automated assessments and human aesthetic evaluations.



**Fig 8.** Correlation between the user evaluations and model predictions (fine-tuning for integrated aesthetics (Q and H))

**Table 5.** Fine-Tuned  $E_Q$  on objective and subjective aesthetics

Metric	MSE	RMSE	PCC
SB	0.03	0.174	0.5
BB	0.01	0.1	0.462
BD	0.002	0.05	0.461
CD	0.007	0.081	0.792
CF	0.012	0.11	0.74
CC	0.003	0.053	0.708



rate consistency	0.045	0.192	0.772
rate clarity	0.043	0.191	0.787
satisfaction	0.038	0.188	0.838

## 5. Discussion

Assessing the aesthetics of a website is a challenging task due to the subjective nature of human perception. Different people may evaluate the aesthetics of the same website differently. While existing approaches use deep learning to evaluate the aesthetics of website design, we advance the current state-of-the-art by integrating computational metrics and human evaluations to improve the evaluation of websites based on various objective and subjective characteristics. Furthermore, our approach uses multi-output regression to evaluate different aspects of website aesthetics simultaneously. The evaluation results demonstrate that was trained accurately to evaluate the objective aesthetics and this due to the high consistent of computational metric in aesthetic scoring, unlike variability human judgment. provides a robust foundation for subsequent fine-tuning processes. As shown in Table 3, fine-tune on subjective aesthetics improves the performance of model cross all subjective aesthetics that are rated by humans, rate consistency, clarity, and satisfaction with correlation 0.931, 0.859, and 0.841, respectively. Inversely, fine-tuning from scratch using ImageNet that results in 0.737, 0.570, and 0.554 for the same subjective aesthetics as shown in Table 4. These results support our hypothesis that the integration of the computational metric and human judgment can improve the deep learning model's performance to evaluate the subjective aesthetic more effectively. Our approach showed an excellent correlation with the human evaluation ( $\rho = 0.88$ ), with the plot in Fig 5 analysis indicating that more than 95% of them agree, i.e., 18 out of the 20 outputs are within the 95% confidence interval. These results outperform other models assessing web design [12], [13], [14]. Regarding performance in literature, our approach obtained an MSE below .04 for all subjective aesthetics, surpassing the assessment of web aesthetics (MSE = .042) Dou et al. [13]. While our approach shows a marginal improvement over Xing et al.'s GUI design assessment (MSE = 0.0222) [10], ours excels in evaluating multiple aesthetics simultaneously through multi-output regression. We agree with Dou et al. [13] that formulating the problem as a regression task is a significant factor in improving that performance as classification models also yielded lower performance results [15]. Our findings draw important conclusions about how to incorporate objective and subjective evaluations in deep learning-based aesthetics

assessment. First off, training models with objective metrics, such as those from Questim, can give them a strong base and help them efficiently capture basic aesthetic features. This method is consistent with earlier studies that highlight the value of reliable and automatic image annotation in the training of models [30]. Second, the human expert evaluation of aesthetics during the fine-tuning process demonstrates how subjective metrics complement each other to enhance the performance of deep learning models. Given the substantial increase in PCC values, it appears that adding human experience can improve the model's capacity to identify fine-grained aesthetic distinctions that objective metrics might miss. However, it is important to acknowledge the limitations of the study. Although the Questim-based model performed well, it may not fully capture the complexity and subjectivity inherent in human aesthetic judgments. The fine-tuned model's higher error rates on certain attributes highlight the ongoing challenges in translating subjective perceptions into quantifiable metrics. It is important to avoid over-generalization of results as the effectiveness of integrating objective and subjective assessments may vary across domains and datasets. Future research should explore more sophisticated methods for integrating different data sources to improve the robustness and applicability of deep learning models in aesthetic evaluation in the future.

## 6. Conclusion

In this paper, we present a novel deep learning-based approach that combines computational metrics and human evaluation to evaluate the visual design aesthetics of webpage. Our approach leverages human judgments and the consistency of computational metrics to overcome the inherent difficulties of subjective aesthetic judgment. The results showed that fine-tuning a base model trained with objective metrics from computational methods was more effective than starting from scratch with an ImageNet-trained model. This was reflected in lower MSE and higher PCC across several subjective aspects, including rating consistency, clarity, and satisfaction. Specifically, the high correlation values were 0.931 for consistency, 0.859 for clarity, and 0.841 for satisfaction. These findings indicate that integrate computation metric and human evaluation can improve the performance of deep learning models in subjective aesthetic of webpage's design. However, our approach was evaluated only on three subjective aesthetic. Future studies can evaluate more subjective aesthetics, different computation metrics with different structure of models. Furthermore, applying this method to other datasets and domains may yield insightful results and enhance the robustness and generalizability of deep learning models in aesthetic assessment.

## Acknowledgements

I would like to express my deep gratitude to everyone who

contributed to this research, whether through advice, technical assistance, or providing resources and references.

## References

- [1] Sutcliffe, *Designing for user engagement: Aesthetic and attractive user interfaces*. Springer Nature, 2022.
- [2] M. T. Thielsch, I. Blotenberg, and R. Jaron, "User evaluation of websites: From first impression to recommendation," *Interact. Comput.*, vol. 26, no. 1, pp. 89–102, 2014.
- [3] T. Lavie and N. Tractinsky, "Assessing dimensions of perceived visual aesthetics of web sites," *Int. J. Hum.-Comput. Stud.*, vol. 60, no. 3, pp. 269–298, 2004.
- [4] S. E. Palmer, K. B. Schloss, and J. Sammartino, "Visual aesthetics and human preference," *Annu. Rev. Psychol.*, vol. 64, pp. 77–107, 2013.
- [5] M. Ramezani Nia and S. Shokouhyar, "Analyzing the effects of visual aesthetic of Web pages on users' responses in online retailing using the VisAWI method," *J. Res. Interact. Mark.*, vol. 14, no. 4, pp. 357–389, 2020.
- [6] L. de S. Lima and C. Gresse von Wangenheim, "Assessing the visual esthetics of user interfaces: A ten-year systematic mapping," *Int. J. Human-Computer Interact.*, vol. 38, no. 2, pp. 144–164, 2022.
- [7] G. Lindgaard, G. Fernandes, C. Dudek, and J. Brown, "Attention web designers: You have 50 milliseconds to make a good first impression!," *Behav. Inf. Technol.*, vol. 25, no. 2, pp. 115–126, 2006.
- [8] J. McCormack and A. Lomas, "Deep learning of individual aesthetics," *Neural Comput. Appl.*, vol. 33, no. 1, pp. 3–17, 2021.
- [9] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rating image aesthetics using deep learning," *IEEE Trans. Multimed.*, vol. 17, no. 11, pp. 2021–2034, 2015.
- [10] Xing, H. Cao, L. Shi, H. Si, and L. Zhao, "AI-driven user aesthetics preference prediction for UI layouts via deep convolutional neural networks," *Cogn. Comput. Syst.*, vol. 4, no. 3, pp. 250–264, 2022.
- [11] J. Zhang, Y. Miao, and J. Yu, "A comprehensive survey on computational aesthetic evaluation of visual art images: Metrics and challenges," *IEEE Access*, vol. 9, pp. 77164–77187, 2021.
- [12] Delitzas, K. C. Chatzidimitriou, and A. L. Symeonidis, "Calista: A deep learning-based system for understanding and evaluating website aesthetics," *Int. J. Hum.-Comput. Stud.*, vol. 175, p. 103019, 2023.
- [13] Q. Dou, X. S. Zheng, T. Sun, and P.-A. Heng, "Webthetics: quantifying webpage aesthetics with deep learning," *Int. J. Hum.-Comput. Stud.*, vol. 124, pp. 56–66, 2019.
- [14] M. G. Khani, M. R. Mazinani, M. Fayyaz, and M. Hoseini, "A novel approach for website aesthetic evaluation based on convolutional neural networks," in *2016 Second International Conference on Web Research (ICWR)*, IEEE, 2016, pp. 48–53.
- [15] L. de Souza Lima, C. G. von Wangenheim, O. P. Martins, A. von Wangenheim, J. C. Hauck, and A. F. Borgatto, "A Deep Learning Model for the Assessment of the Visual Aesthetics of Mobile User Interfaces," *J. Braz. Comput. Soc.*, vol. 30, no. 1, pp. 102–115, 2024.
- [16] M. Zen and J. Vanderdonckt, "Towards an evaluation of graphical user interfaces aesthetics based on metrics," in *2014 IEEE Eighth International Conference on Research Challenges in Information Science (RCIS)*, IEEE, 2014, pp. 1–12.
- [17] M. Xie, S. Feng, Z. Xing, J. Chen, and C. Chen, "UIED: a hybrid tool for GUI element detection," in *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2020, pp. 1655–1659.
- [18] Miniukovich and A. De Angeli, "Computation of interface aesthetics," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 1163–1172.
- [19] M. Zen, N. Burny, and J. Vanderdonckt, "A Quality Model-based Approach for Measuring User Interface Aesthetics with Grace," *Proc. ACM Hum.-Comput. Interact.*, vol. 7, no. EICS, pp. 1–47, 2023.
- [20] P. Koutsabasis and T. G. Istikopoulou, "Perceived website aesthetics by users and designers: implications for evaluation practice," *Int. J. Technol. Hum. Interact. IJTHI*, vol. 9, no. 2, pp. 39–52, 2013.
- [21] Pappas, K. Sharma, P. Mikalef, and M. Giannakos, "Visual aesthetics of E-commerce websites: An eye-tracking approach," 2018.
- [22] S. Eisbach, F. Daus, M. T. Thielsch, M. Böhmer, and G. Hertel, "Predicting rating distributions of Website aesthetics with deep learning for AI-based research," *ACM Trans. Comput.-Hum. Interact.*, vol. 30, no. 3, pp. 1–28, 2023.
- [23] Xing, H. Si, J. Chen, M. Ye, and L. Shi, "Computational model for predicting user aesthetic preference for GUI using DCNNs," *CCF Trans. Pervasive Comput. Interact.*, vol. 3, pp. 147–169, 2021.
- [24] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in

International conference on machine learning, PMLR, 2019, pp. 6105–6114.

- [25] X. Chen et al., “Application of EfficientNet-B0 and GRU-based deep learning on classifying the colposcopy diagnosis of precancerous cervical lesions,” *Cancer Med.*, vol. 12, no. 7, pp. 8690–8699, Jan. 2023, doi: 10.1002/cam4.5581.
- [26] S. Albert et al., “Comparison of Image Normalization Methods for Multi-Site Deep Learning,” *Appl. Sci.*, vol. 13, no. 15, p. 8923, 2023.
- [27] Akça and Ö. Ö. Tanrıöver, “A Deep Transfer Learning Based Visual Complexity Evaluation Approach to Mobile User Interfaces,” *Trait. SignalTS Trait. Signal*, vol. 39, no. 5, pp. 1545–1556, Nov. 2022, doi: 10.18280/ts.390511.
- [28] G. Bonett and T. A. Wright, “Sample size requirements for estimating Pearson, Kendall and Spearman correlations,” *Psychometrika*, vol. 65, pp. 23–28, 2000.
- [29] T. D. V. Swinscow, M. J. Campbell, and others, *Statistics at square one*. Bmj London, 2002.
- [30] M. M. Adnan, M. S. M. Rahim, A. Rehman, Z. Mehmood, T. Saba, and R. A. Naqvi, “Automatic image annotation based on deep learning models: a systematic review and future challenges,” *IEEE Access*, vol. 9, pp. 50253–50264, 2021.