International
Design and
Art Journal

IDA: International Design and Art Journal
ISSN: 2687-5373
Volume: 7, Issue: 2 / 2025

Research Article

Evaluating the aesthetic quality in computer-generated renderings via a comparative analysis

Mustafa Koç[1*], Asst. Prof. Dr. İmdat As[2]

[1]İstanbul Technical University,
Department of Informatics,
Architectural Design Computing,
Istanbul, Türkiye.
kocm21@itu.edu.tr

[2]İstanbul Technical University,
Department of Architecture,
Architectural Design Computing,
İstanbul, Türkiye.
ias@itu.edu.tr

*Corresponding Author

Received: 25.11.2024
Accepted: 03.11.2025

Citation:
Koç, M., As, İ. (2025). Evaluating the aesthetic quality in computer-generated renderings via a comparative analysis. *IDA: International Design and Art Journal, 7*(2), 256-268.

Abstract
In architectural competitions, patrons often receive a multitude of design submissions. Often, it is hard for reviewers to evaluate all submissions in a fair and balanced manner. The research aims to investigate how computational models can assess aesthetic quality in architectural renderings by comparing human-judged scores with algorithmic predictions. Using a dataset of crowdsourced architectural competition entries from Arcbazar, different deep learning models are trained to predict and compare aesthetic scores and generate attribute-based heatmaps. These heatmaps visualize the regions that contribute positively or negatively to the perceived quality of the designs, offering explainable AI outputs. The method includes preprocessing the images, extracting perceptual features, and evaluating model performance through metrics. The results show a high correlation between predicted and actual scores, validating the model's effectiveness. By using machine learning algorithms, a fair and efficient method to assess aesthetics across a large number of submissions is tried to be achieved. This study aims to contribute to the field by providing a transparent and replicable framework for aesthetic evaluation in architecture, bridging human perception and machine analysis. It also demonstrates how explainable AI tools can support assessments in design competitions and stimulate critical dialogue on aesthetics in computational design processes.

Keywords: Deep learning, Aesthetic, Architecture

Extended Abstract

Introduction: Architectural competition platforms like Arcbazar attract a high volume of design submissions, making it challenging for patrons to conduct fair and merit-based assessments. The heavy use of digital renderings has made the proliferation of consistent, objective evaluation processes even more crucial. Traditional evaluation methods are subjective and time-consuming, often leading to biases and inconsistencies. This study explores how machine learning (ML) can automate the aesthetic evaluation of architectural renderings, ensuring fairness and reproducibility. It probes two state-of-the-art ML models, Neural Image Assessment (NIMA) and DeepPhotoAesthetic, regarding their ability to predict aesthetic scores for architectural renderings. The use of ML by this research, hence, tries to fill the gap between human subjective judgment and objective, scalable assessment systems.

Purpose and scope: The primary objective of this study is to develop a tool capable of evaluating the aesthetic quality of architectural renderings submitted in design competitions. The study addresses the application of ML algorithms in effective and efficient ways to determine the architectural aesthetics. Besides, it identifies the relative strengths and weaknesses of the chosen two ML models: NIMA and DeepPhotoAesthetic. These questions arise because of the need for automated systems that can handle the large volume of competition entries while offering reliable aesthetic evaluations. This research focuses on data obtained from the Arcbazar platform, which includes user-provided scores on renderings evaluated for their aesthetic appeal. By processing this data, the study aims to design a system that not only mirrors human judgment but also enhances efficiency, consistency, and fairness in architectural competitions. The outcomes are expected to benefit both project patrons and designers, enabling more informed and equitable decisions.

Method: A mixed-method approach was employed to integrate data preprocessing, model training, and comparative analysis. The dataset used consisted of thousands of architectural renderings submitted to Arcbazar, evaluated on criteria such as aesthetics, idea, function, buildability, and graphics. For this study, the primary focus was on the aesthetics criterion, which formed the basis for training and evaluating the ML models. The renderings underwent preprocessing to standardize resolution and aspect ratio, ensuring compatibility with the requirements of ML models. Images were resized to 1024x512 pixels with a 2:1 aspect ratio, and only high-quality, photo-realistic images were included. This preprocessing step was critical to maintaining consistency and optimizing model performance. The NIMA model, a convolutional neural network (CNN), was retrained using aesthetic scores from the Arcbazar dataset. NIMA's design enabled it to predict overall aesthetic scores by analyzing visual patterns in the renderings. DeepPhotoAesthetic, initially developed for photographic aesthetics, was fine-tuned to evaluate architectural renderings. This model provided attribute-based heatmaps that visually highlighted areas of the image contributing to its aesthetic appeal, offering deeper insights into the evaluation process. Evaluation metrics were employed to measure the accuracy and reliability of the models. Mean Squared Error (MSE) quantified the precision of predictions against actual scores, while correlation coefficients assessed the alignment between predicted and observed trends. Additionally, visual tools such as heatmaps and scatter plots were used to interpret and analyze the results. The performance of both models was systematically compared to determine their effectiveness and limitations in the context of architectural aesthetics. Adjustments were made during the training process to address challenges such as skewed predictions and insufficient data representation.

Findings and conclusion: The study's findings highlight the potential of ML to transform the evaluation of architectural aesthetics. The NIMA model demonstrated effectiveness in predicting overall aesthetic scores but tended to cluster predictions around mean values, thereby limiting its ability to capture the full spectrum of aesthetic diversity present in the dataset. This limitation indicated the need for more nuanced approaches to aesthetic evaluation. DeepPhotoAesthetic outperformed NIMA in attribute-specific evaluations, offering a more detailed and nuanced analysis. Its attribute-based heatmaps were particularly effective in identifying specific visual elements contributing to an image's aesthetic quality. By refining the model to align with architectural criteria, predictions became more consistent with human evaluations. This system not only provides an objective and reproducible method for evaluating design submissions but also offers valuable feedback to designers, helping them refine their work before submission. The findings underscore the importance of tailoring ML models to the unique requirements of architectural applications. This research contributes significantly to the field of computational aesthetics by demonstrating how ML techniques can be applied to architectural design. The study highlights the feasibility of automating aesthetic evaluations, providing a scalable and efficient solution for design competitions. In addition to streamlining the evaluation process, the research facilitates collaboration between designers and patrons by offering data-driven insights into aesthetic quality. The proposed system sets a precedent for integrating AI-driven tools in creative disciplines, enhancing the overall quality and consistency of assessments. Future research directions include expanding the dataset to encompass a broader range of architectural styles and contexts. A larger and more diverse dataset would enhance the generalizability and robustness of the models. Additionally, hybrid approaches that combine the strengths of NIMA and DeepPhotoAesthetic could provide more comprehensive evaluations. Developing real-time feedback tools for designers represents another promising avenue, enabling iterative improvements during the design process. By addressing these areas, subsequent studies can further advance the application of AI in architectural aesthetics, fostering innovation and elevating the standards of evaluation in digital renderings.

Keywords: Deep learning, Aesthetic, Architecture

INTRODUCTION

Online architectural crowdsourcing platforms like Arcbazar attract a high volume of design submissions for each competition, often making it difficult for project-owners to conduct a fair and merit-based assessment. The potential of ML to automate evaluations to develop a tool that can assist project-owners in selecting top designs, i.e., streamlining the evaluation process by filtering and ranking submissions based on aesthetic criteria, thereby presenting project-owners with a curated selection of the most aesthetically compelling designs, is investigated in this study. An ML model that can analyze aesthetic qualities in renderings and predict scores that mirror human evaluation is proposed. It is intended in this research to develop a tool that can be helpful for project owners and designers alike by offering a reproducible method for assessing and improving the aesthetic appeal of design submissions. First, an overview of available tools and methods for aesthetic assessment is given, followed by a deeper discussion of two machine learning models: Neural Image Assessment (NIMA) and DeepPhotoAesthetic. The performance of both models is analyzed in more realistic settings through the execution of an insightful case study using Arcbazar's evaluation system.

NIMA was developed to score hotel rooms, hence it uses a convolutional neural network (CNN) that was trained with datasets such as the AVA dataset for aesthetics and the TID2013 for image quality. While its application was originally dealing with hotel rooms, the structure of NIMA is such that it could also potentially score architectural renderings. DeepPhotoAesthetic is another model developed for general photography scoring. It scores images based on photographic principles such as color harmony, depth of field, and emphasis on objects. The attribute-based heatmap that it provides gives visualizations for its scores, which enable in-depth analysis of different aspects of an image that collectively contribute toward the aesthetic score. For testing these two models within the architectural design discipline, the Arcbazar platform is used. Manual evaluation in the Arcbazar evaluation system is possible for projects on five criteria: Idea, aesthetics, function, buildability, and graphics. In this paper, only the aesthetic criterion is focused. Using data from Arcbazar allowed us to measure how well the NIMA and DeepPhotoAesthetic models predicted aesthetic scores compared to human ratings. Lastly, the results, discussion on the model effectiveness, and possible improvements are presented.

Background

The term "aesthetic" originates from the Greek word "aisthitiki," which refers to knowledge by the senses. Aesthetic preference is subjective since it refers to the likes of a person rather than the characteristics of an object. To the philosopher David Hume, the diversity of interpretation of artwork illustrates the subjective nature of aesthetic judgment. This subjectivity suggests that aesthetic perceptions are determined by cultural, personal, and experiential factors. Still, despite these differences, some universal elements allow for shared perceptions of beauty, which form the basis for shared aesthetic experience (Zangwill, 2001: 122). Aesthetic perception depends on experience and familiarity; for instance, initial indifference to musical patterns can lead to enjoyment when continuously exposed to the same soundtrack, because familiarity strengthens cognitive associations (Hoenig, 2005: 14). Abraham Moles developed the theory of aesthetic perception further in his information-theoretical model, which postulates that memory and perceived redundancy play an important part in the aesthetic experience. Moles went on to devise the term "differential information," a component of aesthetic value that depends partly on an object's novelty and its relation to an individual's cumulative memory of similar stimuli. These considerations raise questions as to the extent to which such dynamics might be analytically and quantitatively transferable by ML. While aesthetic perception is often considered elusive, new perspectives on understanding and predicting human aesthetic judgments based on quantifiable features within a given dataset surface with the advance of ML.

Artificial Intelligence (AI) is used in many stages of a building's life-cycle, from generative systems that create conceptual designs (As et al. 2019: 433), to systems that develop landscape designs (Senem et al., 2023: 8), to models that explore composing entire cities. Presently, it is observed that advanced computing systems start to emulate also certain aspects of human cognition, such as aesthetic decision-making (McCormack & Lomas, 2020: 9). Aesthetic computing, a term introduced by Fishwick (2006), brings the theory of art into computing by taking into consideration both the analytical and creative dimensions of aesthetics. For instance, with ML as its complement, analytic aesthetics will make it possible for computers to process voluminous complex data sets, thus coming up with patterns and relationships that may bring insight into aesthetic judgment. The two interdisciplinary fields, computational aesthetics and aesthetic computing, bridge digital technology with fine arts, design, computer science, cognitive science, and philosophy and define the subject area of computational aesthetics to be developing algorithms that can independently determine aesthetic quality in visual expressions and then generate aesthetic and engaging content (Bo et al., 2018: 1). That means, this field researches how ML algorithms can be trained in such a manner to identify and then mimic collective aesthetic preferences, in turn allowing humans and machines to collaborate better in creative contexts (McCormack & Lomas, 2020: 13). By understanding how computational aesthetics can be applied to architectural design entries, it not only automates such assessments but also goes deeper into the insights on how people perceive design elements; it opens exciting new possibilities for creative fields. While AI evolves day by day, systems can already interpret, predict, and even generate aesthetic qualities from their measurable characteristics, closing the gap between human perception and machine analysis (Yu et al., 2020: 1). This study takes the next step toward machine learning techniques that can estimate and predict aesthetic aspects within architectural renderings and expands on translating aesthetic judgment through data-driven methods.

Architecture is a domain unto itself in which aesthetic evaluations leveraging ML come into play with a notable significance. According to Franco (2019: 394), unlike other forms of art, architectural beauty needs professional evaluation. Evaluation of the work of architecture involves functionality and aesthetics that the untrained eye may not be able to decipher. In architectural competitions where aesthetic features may be considered critical, no universal definition of beauty exists. Aesthetic judgements reflect personal and professional biases (As, 2019: 271). This 'spectrum of positions' underlines the importance of developing a systematic method in the architectural design evaluation, especially when hundreds of design entries are to be processed and considered in a competitive environment (Cross, 2001: 47). The latest progress that has been made within the AI space has seen the advent of advanced algorithms, including those like neural networks and decision trees, which can consider very complicated, non-linear data. This is indeed its strength and makes them quite suitable for computational aesthetics (Aydin et al., 2021). Algorithms that are conceptualized to function with ambiguous and multi-layered information afford more precise predictions of complex, multi-dimensional data and allow aesthetic judgments to be generalized across extensive datasets (Basheer & Hajmeer, 2000: 28). By processing aesthetic criteria computationally, such systems offer new possibilities in architecture and enable more consistent and scalable assessments of designs. Online architectural competition platforms, such as Arcbazar, represent an ideal environment in which computational aesthetics can be applied for non-biased and systematic ranking and judging of the designs that will enhance the whole review process.

METHOD

To evaluate the aesthetic quality of architectural renderings, an initial review of five aesthetics assessment algorithms is conducted: the Expedia-NVIDIA collaborative model, RAPID, NIMA (Neural Image Assessment), the Universal Image Attractiveness Ranking Framework, and DeepPhotoAesthetic (Table 1).

Table1. Reviewed aesthetic assessment algorithms

ALGORITHM	PURPOSE	ADVANTAGES	DISADVANTAGES
Expedia-Nvidia Collaborative Model	Developed to assess the aesthetic quality of hotel room images for Expedia's evaluation system.	Specifically trained on interior images, providing focused evaluations for indoor spaces.	Limited to hotel room aesthetics, may not generalize well to other types of architecture or renderings.
RAPID	A general-purpose aesthetic ranking model that rates image quality and attractiveness.	Offers a broad evaluation of image attractiveness, potentially adaptable to different contexts.	Lacks specialization for architectural or design-focused images; may produce generic aesthetic scores.
NIMA	Predicts aesthetic scores by correlating with human ratings, using a CNN to assess image quality.	Can predict both overall aesthetic scores and score distributions, providing insights into public perception.	Requires large datasets for accurate predictions; may struggle with architectural renderings not in training set.
Universal Image Attractiveness Ranking Framework	Ranks image attractiveness based on universally appealing features across image types.	Broad applicability; aims to identify universally attractive features, making it flexible across image genres.	Generalized attractiveness assessment may overlook specific aesthetic elements important in architecture.
DeepPhotoAesthetic	Assesses aesthetic quality using photographic principles, with attribute-based heatmaps.	Provides attribute-based heatmaps, enabling a detailed analysis of specific aesthetic factors within an image.	Initially designed for photographs, so may require adjustments for architectural renderings.

Out of these, NIMA and DeepPhotoAesthetic are selected for a more detailed exploration, as their outcome seemed more promising for this work. NIMA was originally developed for assessing the aesthetic quality of hotel rooms and thus had been trained on interior room shots. Building upon an earlier model at Google called the Neural Image Assessment model, later further adapted in a collaboration between Expedia and Nvidia, NIMA evaluates images along two key dimensions: aesthetic quality and technical quality (Lennan & Tran, 2018). It is first trained on the AVA dataset (a large open-source audio-visual database for aesthetic analysis) and then fine-tuned for image quality using the TID2013 dataset (containing 25 reference images and 3000 distorted images), to perform aesthetic evaluation on architectural work. Meanwhile, DeepPhotoAesthetic was designed for aesthetic evaluation in photographs, utilizing a dataset of images rated by users. This model is inspired by "Learning Photography Aesthetics with Deep CNNs" (Malu et al., 2017: 3). It has been modified to work with eight attributes, namely, balancing elements, content, color harmony, depth of field, light, object emphasis, rule of thirds, and vivid color, to form an aesthetic framework (Kong et al., 2016: 8). With these features, the model is allowed to give aesthetic quality evaluation using attribute-based heatmaps, visualizing high and low aesthetic scores over image-specific regions. In this study, the original photographic attributes are replaced with architecture-specific criteria, including beauty, function, buildability, graphics, and idea, in aligning the model toward architectural competitions.

For this study, the following programming environments are used: Python, PyTorch, and PyCharm. Python is selected because of its comprehensive libraries for machine learning, data processing, and image manipulation. Pytorch is adopted to implement the NIMA and DeepPhotoAesthetic models. Its flexibility allows us to fine-tune the model with high efficiency. Finally, PyCharm is used as the main integrated development environment (IDE) to facilitate code organization and debugging. It was especially helpful in the pre-processing of data and in the training of models. Among the evaluation tools, the mean squared error (MSE), correlation coefficient (R^2), and attribute-based heatmaps are used. The MSE is used for calculating the precision of the predicted aesthetic scores concerning the actual scores provided by Arcbazar users. This metric quantifies the average squared differences between predicted and actual values; thus, it allows for a precise assessment of model performance. R^2 is used to find out the level of similarity in the scores between predicted ones and actual ones, hence estimating how well these models captured the aesthetic trends present within the user evaluations. Finally, attribute-based heatmaps are used to visualize those regions within an image that contribute to a higher or lower aesthetic score. These heatmaps were most useful in understanding which attributes affected the aesthetic evaluations for architectural work. In this context, the statistical tools used in this study were descriptive statistics and comparative analysis (t-test). Descriptive statistics entails basic statistical analyses, such as the calculation of means, medians, and standard deviations. It is applied to both the predicted and actual scores to summarize the distribution of scores and find patterns in the model's predictions. Then a t-test is run on the differences of NIMA and DeepPhotoAesthetic's predicted scores against Arcbazar users' true scores to see for each model whether the differences between the pairs of scores were statistically significant. This will let us judge the precision with which the models reflect human aesthetic judgments.

The visualization tools were Matplotlib, Seaborn, and attribute activation maps. Matplotlib is a very handy library to plot model performance metrics such as the prediction vs. actual score distribution and training loss trends. This made it possible to create comparative visualizations, which shed light upon the consistency of the model's predictions. On the other hand, advanced statistical visualizations are powered by Seaborn. Complementing Matplotlib, it offered more intuitive correlation heat maps, where it could be easier to see relationships between predicted and actual scores and to get an idea of the overall spread of data points. And finally, attribute activation maps are used to visually highlight image regions associated with higher aesthetic scores. These maps helped interpret the model's focus areas, which basically underlined the understanding of architectural images of varied visual impact across different regions. The main cloud computing and hardware resources were Google Colab and NVIDIA GPUs. All model training is done on Google Colab using its cloud-based environment to avail the power of NVIDIA GPUs to speed up the training processes. This enabled larger batch processing and provided enough computational resources to fine-tune such large models as in DeepPhotoAesthetic, which needs heavy processing because it does attribute-based analysis of architectural renderings.

Scoring benchmarks, e.g., scores received from project-owners, designers, and experts through crowdsourced data received from Arcbazar, are prepared to be able to compare the automated results with human evaluations. These scores gave a benchmark against which ML models were trained, and their output was evaluated. This has allowed the generalization of aesthetic judgment because the diversity of user feedback allows for the prediction of a wide range of scoring preferences in architectural competitions. Figure 1 illustrates the workflow diagram for this study.

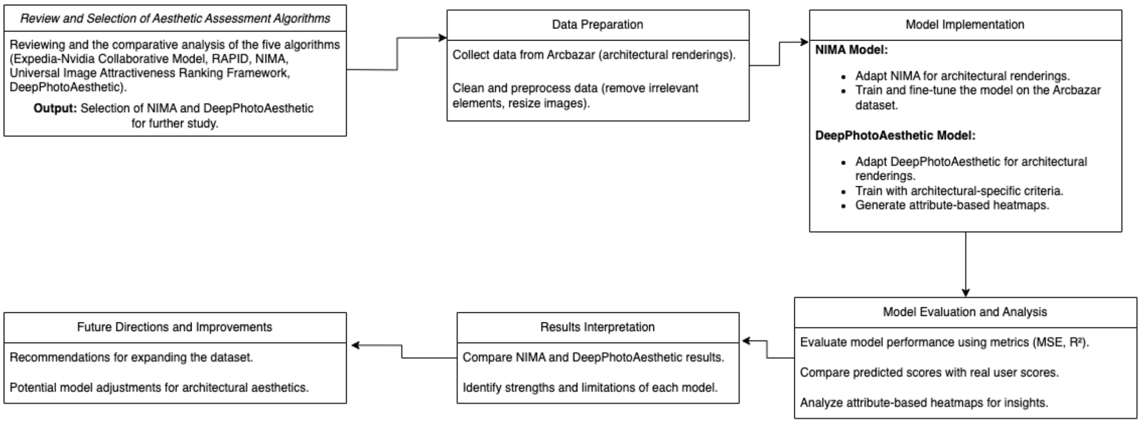

Review and Selection of Aesthetic Assessment Algorithms
Reviewing and the comparative analysis of the five algorithms (Expedia-Nvidia Collaborative Model, RAPID, NIMA, Universal Image Attractiveness Ranking Framework, DeepPhotoAesthetic).
Output: Selection of NIMA and DeepPhotoAesthetic for further study.
Data Preparation
Collect data from Arcbazar (architectural renderings).
Clean and preprocess data (remove irrelevant elements, resize images).
Model Implementation
NIMA Model:
Adapt NIMA for architectural renderings.
Train and fine-tune the model on the Arcbazar dataset.
DeepPhotoAesthetic Model:
Adapt DeepPhotoAesthetic for architectural renderings.
Train with architectural-specific criteria.
Generate attribute-based heatmaps.
Model Evaluation and Analysis
Evaluate model performance using metrics (MSE, R²).
Compare predicted scores with real user scores.
Analyze attribute-based heatmaps for insights.
Results Interpretation
Compare NIMA and DeepPhotoAesthetic results.
Identify strengths and limitations of each model.
Future Directions and Improvements
Recommendations for expanding the dataset.
Potential model adjustments for architectural aesthetics.

Figure 1. Flowchart of the workflow

Case Study: Evaluating competition entries at Arcbazar

The evaluation of design entries to architectural competitions has major challenges, most importantly, the sheer number of submissions per competition and the subjectivity in human assessments. For example, the Opera National de Paris building competition in 1983 received 756 design entries (As et al., 2019). These large numbers are not uncommon for many popular design competitions. It is humanly impossible for a limited number of jurors to assess all submissions consistently and fairly. Also, aesthetic judgment is subjective by nature, and variations between reviewers could lead to “unfair” decisions. For this purpose, developing an ML system that can automate some of the ratings and rankings of projects is aimed. For this purpose, ML models were trained on large amounts of architectural images, with the hypothesis that they would be capable of automating aesthetic scores for architectural projects. This may enable a more objective ranking process. Moreover, the system would not only help to rank vast numbers of design submissions but also could give valuable feedback to designers about their own work before the submission deadline and thus empower them to refine their work before their final submissions.

In this study, the evaluation scores from different user groups on Arcbazar, including project-owners, experts, and general users, are considered. Submissions were evaluated across five criteria: Idea, aesthetics, function, buildability, and graphics (As & Nagakura, 2016: 71). The analysis specifically focused on the aesthetics criterion to develop a targeted aesthetic assessment model. Users, who include project-owners, experts, and general users, on the Arcbazar system rate the images of the designs presented. These scores are on a scale of 1-10 and treated as ground truth for model training. The scores are normalized for the ML models to ensure standardization. The data received from Arcbazar is preprocessed and standardized for the use of ML models. This ensured that the data used for training and evaluation was uniform. First, the NIMA model is trained with a subset of renderings that are aligned with the Arcbazar scoring criteria. Then, the DeepPhotoAesthetic is tested with its pre-trained model and fine-tuned with the interior and exterior architectural images from the dataset to align them with the aesthetic standards of Arcbazar. Each of the ML models was tailored to Arcbazar’s criteria and targeted only aesthetics. While NIMA produced only scoring predictions, DeepPhotoAesthetic provided further detail via attribute activation maps, which visually indicated how aesthetic scores were distributed across different regions of an image. Based on the performance of these models, a prototype system that automatically evaluates and ranks renderings submitted on Arcbazar has been developed. This case study demonstrates the potential of ML-driven aesthetic evaluation in online architectural competitions, providing a scalable solution to challenges associated with high submission volumes and subjective judgment. By implementing and fine-tuning NIMA and DeepPhotoAesthetic for architectural renderings, how ML can contribute to fairer and more systematic aesthetic assessment processes is illustrated.

Testing NIMA

The NIMA system, developed by Google, is a pre-trained model on millions of hotel images rated by a diverse population of regular hotel-goers worldwide. NIMA uses a convolutional neural network (CNN) to assess images based on predicted aesthetic scores, with the objective of achieving a high correlation with human ratings rather than a binary classification of low or high scores (Talebi & Milanfar, 2018). One of the strong points of NIMA is its flexibility-it can be trained on datasets that capture both aesthetic and pixel-level qualities, thus finding wide applications in aesthetic evaluation tasks. In the present work, NIMA is adapted and retrained using the dataset comprising architectural renderings scored with user evaluations from Arcbazar. This retraining was iteratively done to bring NIMA's predictive capability in line with the specific aesthetic criteria of this study. The re-trained NIMA model tended to predict scores close to the average aesthetic ratings of the repository without extreme deviations, thus aligning with the general aesthetic consensus in the dataset.

The image dataset had to undergo preprocessing to make it fit the input requirements of NIMA. Each design submission in Arcbazar contains several renderings in various styles and formats, such as monochromatic sketches, colored plans, or hyper-realistic visualizations. In the original experimentation of NIMA, only photograph-like, hyper-realistic renderings and well-composed furnished interior views were included since this more closely aligned with the types of images that the model was originally trained on. High-fidelity architectural images that better reflect aesthetic qualities are selected, and all renderings were standardized regarding resolution (1024x512 pixels), and aspect ratio (2:1) to ensure the quality of the input was consistent. Such pre-processing was an important step in making sure that the model performed optimally in its prediction results. For augmenting and increasing the number of images in the training dataset, techniques like flipping, rotating images, or applying various distortions are used (Figure 2).

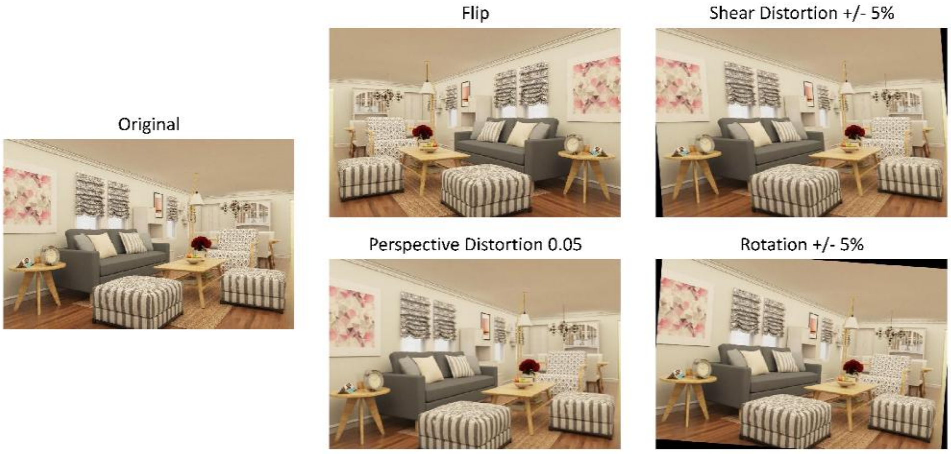


Figure 2. Augmentation: Image transformations

The assumption is that the aesthetic appeal of hotel rooms, rated by NIMA, could translate well to residential settings, assuming aesthetic principles are similar between the interior design of hotels and the design projects on Arcbazar. While testing the "out-of-the-box" NIMA model on architectural renderings from the dataset, an inclination toward mean values is noticed for aesthetic assessment scores; this often led to low scores versus the actual rating in the repository. Figure 3 shows that a comparison of regular users and expert votes indicates regular users tend to vote similarly to experts for extreme scores. However, for less extreme scores, the distribution becomes highly divergent, with regular-user ratings greatly differing from those of experts.

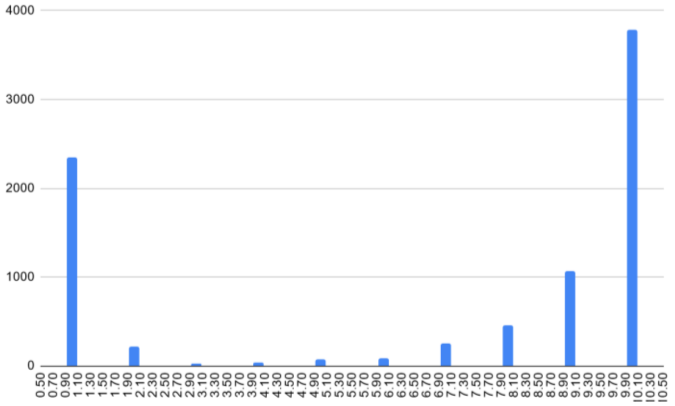


Figure 3. Degree of similarity between designer vs expert votes

Figure 4 presents the aesthetic scores predicted by the NIMA model, indicating a tendency for scores to cluster around the average. This is further elaborated in Figure 5, i.e., the difference between high and low predicted scores is minimal. While NIMA performs well in predicting scores around the mean of the dataset, it does not capture the full distribution of real scores, which range from minimum to maximum values.

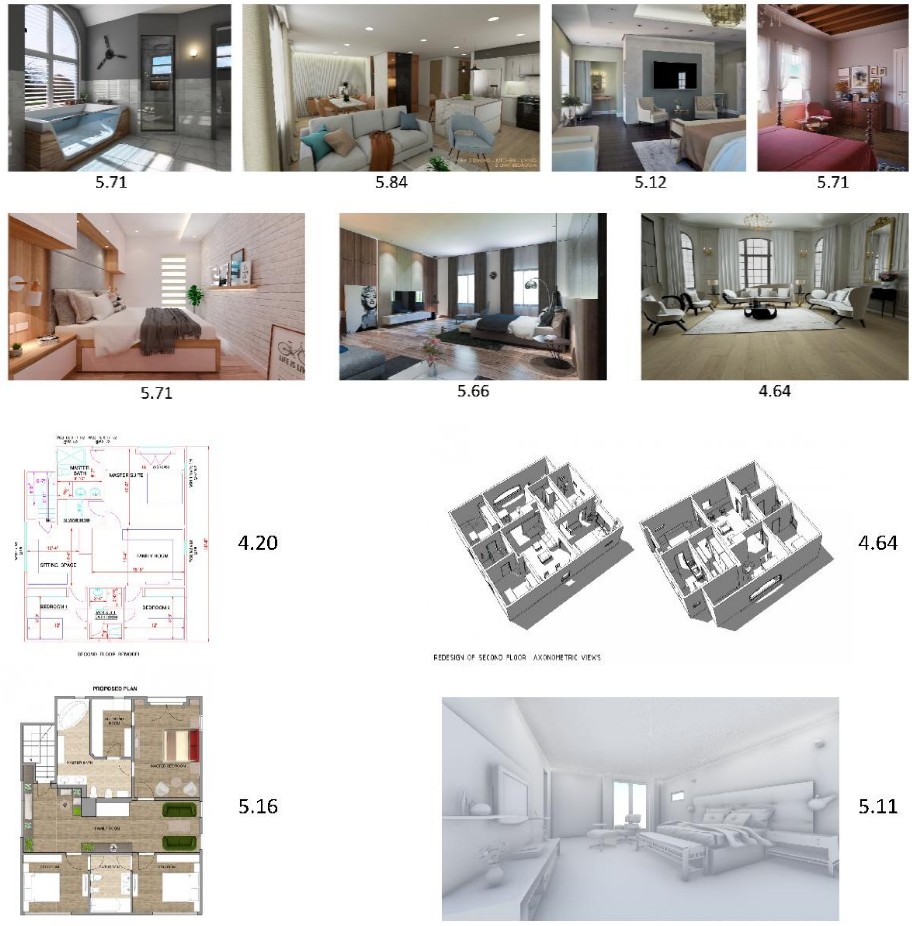


Figure 4. High-Quality Scoring & Low-Quality Scoring

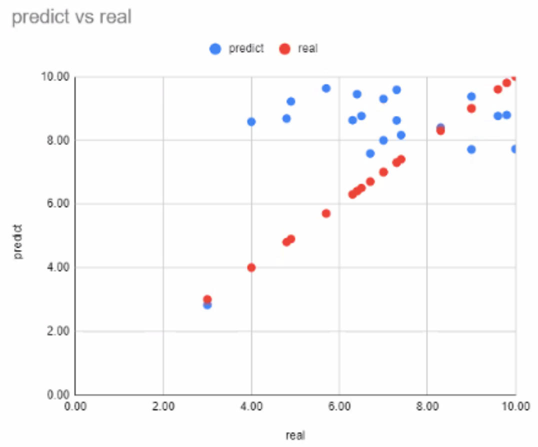


Figure 5. Nima score predictions

Therefore, this model could not predict aesthetic criteria at a satisfactory level to evaluate the characteristics of architectural work, since its predictions could not represent diversity and range as real human evaluations do. With these limitations, the DeepPhotoAesthetic is selected for testing to get closer to a more nuanced approach for aesthetic assessment.

Testing DeepPhotoAesthetics

The existing computational models for aesthetic evaluation typically focus on a single aesthetic score or class, and do not provide any insight into the specific attributes that add up to a particular image's quality. DeepPhotoAesthetic attempts to overcome this limitation by predicting an overall aesthetic score with human-interpretable explanations based on multiple aesthetic aspects (Malu et.al., 2017: 3). The model leverages a multi-task deep convolutional neural network (DCNN) that learns simultaneously multiple aesthetic criteria along with the total score, hence enhancing accuracy and interpretability. It is trained to evaluate images based on eight attributes commonly associated with photographic quality: Balancing elements, content, color harmony, depth-of-field, light, object emphasis, rule of thirds, and vivid color. In the first iteration of the test, a few elements, such as balancing elements, light, rule of thirds, had minimal impact on the overall aesthetic score for architectural renderings. As a result, these attributes were eliminated from the evaluation to divert this model's focus on characteristics that are closer to architecture.

Within the model, heatmaps are utilized as a form of visual interpretability to determine which areas of an image contribute most to the model's predicted aesthetic score. Heatmaps are generated through a process that involves gradient-based class activation mapping (Grad-CAM). After the image passes through the convolutional layers, gradients of the output (aesthetic score) are computed with respect to the last convolutional layer feature maps. These gradients are then weighted and averaged to produce a coarse localization map indicating important regions. In architectural renderings, this enables a designer or critic to visually understand which spatial components (such as facade articulation, lighting, texture detailing, or perspective depth) the model considers aesthetically significant. The generated heatmaps are upsampled and overlaid on the original renderings, offering intuitive, human-interpretable feedback that bridges the gap between machine prediction and human visual reasoning.

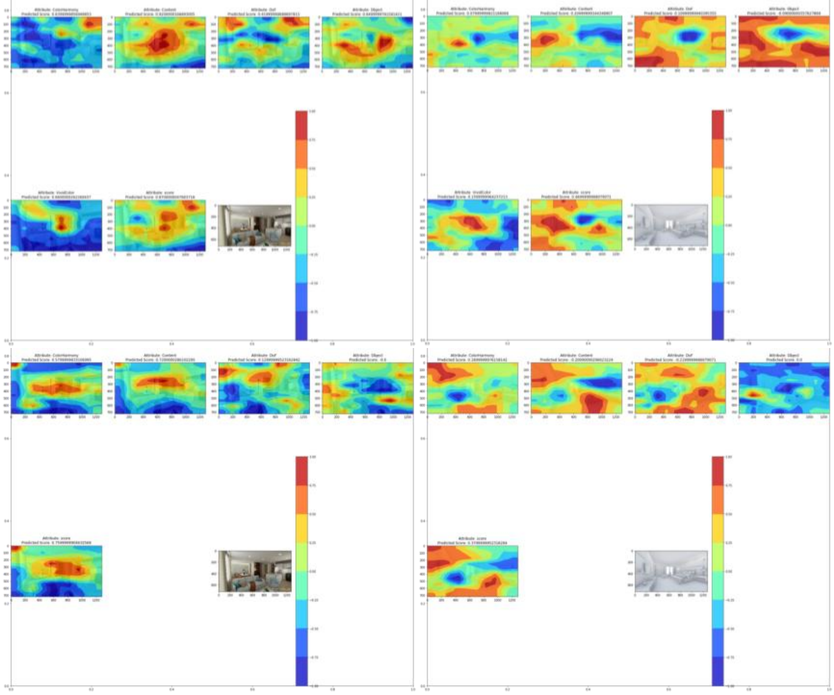


Figure 6. Results from a pre-trained model of DeepPhotoAesthetics with different parameters

Figure 6 shows the output of DeepPhotoAesthetic's pre-trained model on architectural renderings with its own original attributes. Initial results using its default parameters were poor. The results showed that, for example, "vivid color" always showed a low score and contributed little to the model's predictive ability on the data. Thus, this attribute is excluded in a subsequent run, and the algorithm is retrained with the remaining four attributes: content, color harmony, depth-of-field, and object emphasis. This resulted in a sharp increase in improvement within the predicted scores, yielding results that more closely fit within the range of maximum and minimum scores shown in the original data. The refined model displayed a better distribution of aesthetic scores, enhancing its capability to differentiate high-quality renderings from lower-quality ones based on architectural aesthetic criteria.

The DeepPhotoAesthetic model is trained with a dataset of 2,000 architectural renderings, which increases the model's ability for better prediction and evaluation of aesthetic scores. According to the results of the preliminary experiment, the dataset is divided into two major categories: interior and exterior renderings. This separation was done to enhance the accuracy of its prediction, since aesthetic features and visual characteristics differ between interior and exterior images. One of the key advantages of DeepPhotoAesthetic is that it produces heatmaps for individual attributes, thus giving a visual representation of the distribution of aesthetic scores in different regions of an image. This was particularly helpful in the study because it allowed us to tell which parts of the score were most contributing to the overall score (Figure 7).

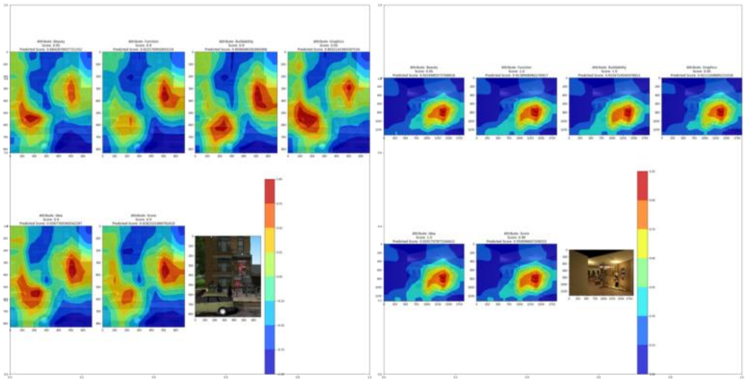


Figure 7. Heatmaps of exterior and interior images with the DeepPhotoAesthetic model

The results from the DeepPhotoAesthetic model outperformed both the NIMA model and the pre-trained DeepPhotoAesthetic model. Yet even with this progress, there was still a significant gap between the real scores and the model's predictions. To further improve the performance, some parameter tuning work is done, such as input shape and optimizer, to adapt the training to the architectural rendering dataset.

Figure 8 shows the process of data preparation. First, the data from the online repository is collected, and the images are cleaned manually using Adobe Photoshop to maintain consistency in their quality. Then, the images are resized to 512x256 pixels using PyCharm, a Python IDE, to decrease training time but without sacrificing important image features. After resizing, packages and parameter settings are modified on PyCharm to fine-tune the algorithm to the architectural renderings. One was scaling the loss function to allow the data's score range of 0-10 and matched it with the -1 to 1 scale of the original DeepPhotoAesthetic. For initial testing, the model is exclusively trained on interior renderings so that results may be directly compared with the NIMA model. Table 2 depicts the outcomes of training for interior renderings using losses that were recorded for every attribute across epochs, indicating a significant reduction over 20 epochs.

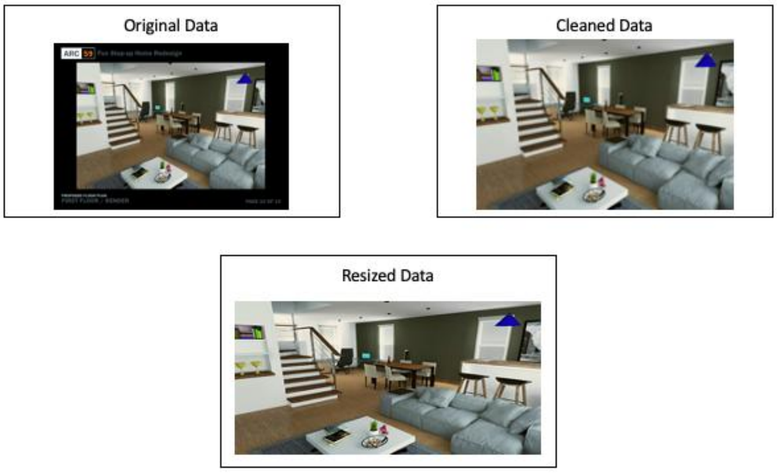


Figure 8. Data preparation

Table 2. Output of training DeepPhotoAesthetic model with interior renderings across 20 epochs

Epoch	Beauty	Function	Buildability	Graphics	Idea	Score	Total Loss
0	0.031129	0.031154	0.030242	0.031635	0.021784	0.043113	0.189056
19	0.011930	0.012033	0.014218	0.011471	0.005830	0.010392	0.065876

The low loss values indicate that adjustments to the algorithm were effective in adapting it for architectural renderings (Figure 9).

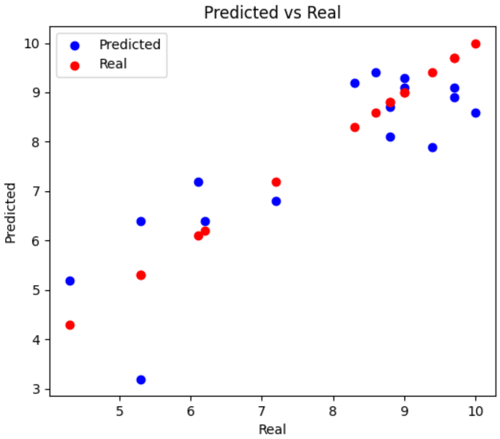


Figure 9. DeepPhotoAesthetic predictions (blue) vs real scores (red)

Although promising, the dataset remains relatively small. However, these results confirm that an adapted ML model can be used to estimate attribute-specific aesthetic scores and overall aesthetic scores.

FINDINGS

Compared to architectural renderings evaluation, aesthetic assessment scores produced by the off-the-shelf NIMA model were always lower and tended towards the mean, i.e., were far from scores obtained from Arcbazar. Scores of NIMA were limited toward the average score and maximum value in the dataset and failed to reflect the whole spectrum of aesthetic scores as observed in real evaluations. On the other hand, the tests of architectural renderings with the DeepPhotoAesthetic model, using its original photographic criteria, yielded more accurate overall score predictions. Additionally, DeepPhotoAesthetic's heatmap feature offered valuable insights into the aesthetic aspects that influenced each image's score. Training DeepPhotoAesthetic with architectural images demonstrated that, even with a relatively small dataset, the model could predict attribute scores closely aligned with actual user evaluations. The NIMA system has certain advantages, such as predicting both overall aesthetic scores and the distribution of human opinion scores. However, the available data for this study limited NIMA's effectiveness, as it was insufficient to fully train the model. Furthermore, to align with NIMA's intended use, the training is restricted to interior images from specific perspectives, which imposed significant limitations. Despite these adjustments, NIMA's predicted scores remained higher than actual scores, and the distribution was constrained to average and maximum scores only, diverging from the true spread of user evaluations. In contrast, DeepPhotoAesthetic operates on a pre-trained model developed for photography, which, although not fully optimized for architectural images, outperformed NIMA in accuracy. The model's ability to separate attribute scores and display them via heatmaps aligned well with the goals of this study. Upon applying DeepPhotoAesthetic to interior images, the predictions came somewhat closer to the real scores compared to NIMA results (Figure 10). Attribute-specific heatmaps intuitively determined aesthetic elements, which made a more in-depth visual analysis of architectural renderings with photographic attributes possible. The more additional training on architectural images was done, the more meaningful the scores as well as heatmaps from the DeepPhotoAesthetic became, empowering the model in its ability to assess the aesthetics of the renderings. This research indeed reveals what attribute-based analysis by DeepPhotoAesthetic can do in terms of assessing the aesthetics of architecture.

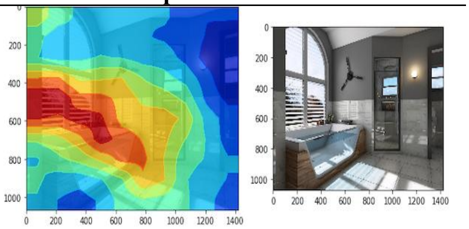
Arcbazar Data	NIMA Result	DeepPhotoAesthetic
		
Real Score: 8	Predicted Score: 5.71	Predicted Score: 8.1

Figure 10. Comparison of the results from NIMA and DeepPhotoAesthetics against the real score

CONCLUSION

In this paper, the challenges of assessing aesthetic qualities in architectural projects submitted to design competitions are explored. A high volume of submissions, together with the inherently subjective nature of human judgment, makes it difficult to conduct fair and consistent evaluations. Therefore, two machine learning models, NIMA and DeepPhotoAesthetic, are tested to train a prototype model that can automatically evaluate and rank renderings with respect to aesthetics. Testing and refinement showed that both models were indeed helpful, but the attribute-based heatmaps from DeepPhotoAesthetic allowed for a finer understanding of aesthetic qualities due to the capture of specific visual attributes relevant to architectural renderings. NIMA, though effective in scoring overall aesthetics, showed limitations in predicting a broader range of scores, since predictions from this model tended to center around mean values. Together, these models have demonstrated how ML can support aesthetic scores to standardize the evaluation process. The work has several limitations in terms of dataset size, data quality, evaluation criteria, and model constraints. The dataset was relatively small, which reduces the model's generalization ability and affects its accuracy. A large and diverse dataset will increase the robustness of the prediction made by the system. The image quality and style variations within the dataset may affect the model performance, the performance of NIMA, for example, is sensitive to resolution and clarity. Moreover, the research focused only on aesthetics, but it could be improved by the incorporation of further criteria such as function and buildability. Moreover, NIMA and DeepPhotoAesthetic were originally designed for general photographic aesthetics rather than architectural renderings, which may limit their effectiveness. Based on these findings, future studies in this area could expand on these shortcomings. Incorporating data from various architectural styles and sources may help improve model accuracy and generalizability. Also, the introduction of new, architecture-specific criteria, such as spatial dynamics or material qualities, may be more congruent with the evaluation of architectural aesthetics.

This paper opens many possibilities for AI to have a more active role in the evaluation and perhaps even help in the process of design production in architectural competitions. It is envisioned that future models will bridge the gap between subjective human judgment and objective machine analysis, contributing to a much fairer, more consistent, and insightful assessment of architectural work. It is foreseen that, in the long run, an AI system can design architectural projects or be used as a tool to provide template designs and alternatives that other designers can continue to work with. The aesthetic values these AI-generated designs will have, how these values are formulated, and what relationships the machine will establish in them are crucial. The ML model developed and tested herein will play a part in that understanding of these values and relationships. A system that can predict the aesthetic value of an architectural project could support such understanding significantly. As Cross (2001: 44) noted, the use of computers to understand human design processes has been done throughout history. While this perspective remains valid, computers could extend beyond this role, evolving into essential tools for perceiving and exploring complex design concepts such as "aesthetics."

Authors' Contributions
The authors contributed equally to the study.

Competing Interests
There is no potential conflict of interest.

Ethics Committee Declaration
The study does not require ethics committee approval.

REFERENCES

Arcbazar.com, Inc. (2022). *Arcbazar*. https://www.arcbazar.com/ (12.09.2023).

As, I., Pal, S., & Basu, P. (2019). Composing Frankensteins: Data-driven design assemblies through graph-based deep neural networks. In *The 107th Annual Meeting BLACK BOX: Articulating Architecture's Core in the Post-Digital Era* (pp. 433-438). ACSA.

As, I. (2019). Competitions in a networked society: Crowdsourcing collective design intelligence. In *Proceedings of the BLACK BOX: Articulating Architecture's Core in the Post-Digital Era* (pp. 268-273). ACSA.

As, I., & Nagakura, T. (2016). Architecture for the crowd by the crowd: A new model for design acquisition. *International Journal of Architecture and Urban Studies, 1*(2), 68-76.

Aydin, Y., Parham, M. A., & Hale, J. (2021). Using machine learning techniques to predict esthetic features of buildings. *Journal of Architectural Engineering, 27*(3), 04021023. https://doi.org/10.1061/(ASCE)AE.1943-5568.0000477

Basheer, I. A., & Hajmeer, M. (2000). Artificial neural networks: Fundamentals, computing, design, and application. *Journal of Microbiological Methods, 43*(1), 3-31. https://doi.org/10.1016/S0167-7012(00)00201-3

Bo, Y., Yu, J., & Zhang, K. (2018). Computational aesthetics and applications. *Visual Computing for Industry, Biomedicine, and Art, 1*(1), 6. https://doi.org/10.1186/s42492-018-0006-1

Talebi, H., & Milanfar, P. (2017, September 15). *NIMA: Neural Image Assessment.* arXiv. http://arxiv.org/abs/1709.05424 (09.11.2024).

Malu, G., Bapi, R. S., & Indurkhya, B. (2017, July 13). Learning photography aesthetics with deep CNNs. *arXiv*. https://arxiv.org/pdf/1707.03981.pdf (15.11.2024).

McCormack, J., & Lomas, A. (2020). Deep learning for individual aesthetics. *Neural Computing and Applications, 32*, 11727-11743. https://doi.org/10.1007/s00521-020-05376-7

Cross, N. (2001). Can a machine design? *Design Issues*, *17*(4), 44-50. http://www.jstor.org/stable/1511919

Franco, A. B. (2019). Our everyday aesthetic evaluations of architecture. *British Journal of Aesthetics, 59*(4), 393-412. https://doi.org/10.1093/aesthj/ayz018

Hoenig, F. (2005). Defining computational aesthetics. In *Computational Aesthetics in Graphics, Visualization and Imaging* (pp. 13-18). The Eurographics Association.

Lennan, C., & Tran, D. (2018, October 30). *Deep learning hotel aesthetics photos. NVIDIA Developer Blog.* https://developer.nvidia.com/blog/deep-learning-hotel-aesthetics-photos/ (12.09.2024).

Kong, S., Shen, X., Lin, Z., Mech, R., & Fowlkes, C. (2016, July 27). *Photo aesthetics ranking network with attributes and content adaptation. arXiv.* https://arxiv.org/pdf/1606.01621.pdf (05.10.2024).

Senem, M. O., Koç, M., Tunçay, H. E., & As, İ. (2023). Using deep learning to generate front and backyards in landscape architecture. *Architecture and Planning Journal (APJ), 28*(3), 1-10.

Zangwill, N. (2001). *The metaphysics of beauty*. Cornell University Press. http://www.jstor.org/stable/10.7591/j.ctv1nhmzk

Yu, Y., Beuret, S., Zeng, D., & Oyama, K. (2018, September 3). Deep learning of human perception in audio event classification. *arXiv*. https://arxiv.org/abs/1809.00502 (07.12.2024).

Authors' Biographies

Mustafa Koç is an engineer who works in inter/trans disciplinary studies. After graduated from Bilkent University, completed his master's degree in both engineering and architecture in TOBB University of Economics and Technology. During his graduate studies, he focused on machine learning and its applications in various disciplines. Currently he is a Ph.D. candidate at İstanbul Technical University in Architectural Design Computing. He is studying machine learning, synthetic data, and decision support systems.

İmdat As is an architect, entrepreneur, academician, and the recipient of the prestigious Turkish Scientific Research Council (TUBITAK) 2232 Grant and Fellowship. He leads the City Development through Design Intelligence (CIDDI) lab at Istanbul Technical University (ITU). His work focuses on new technologies that shape the morphology of the future city. He received his doctoral degree from the Harvard University Graduate School of Design (GSD), his M.Sc. in architecture from the Massachusetts Institute of Technology (MIT), and his B.Arch. from the Middle East Technical University (METU) in Ankara, Turkey. He co-authored, Dynamic Digital Representations in Architecture: Visions in Motion (Taylor & Francis, 2008), and two books on AI in Architecture (Taylor & Francis, 2021 and Elsevier, 2022). He founded Arcbazar.com in 2011- a crowdsourcing platform for architectural design. Arcbazar has been featured as one of the "Top 100 Most Brilliant Companies" by Entrepreneur Magazine.