Megaron, Vol. 21, No. 1, pp. 32–41, March 2026

Megaron
https://megaron.yildiz.edu.tr - https://megaronjournal.com
DOI: https://doi.org/10.14744/megaron.2026.07448

Article

Deep learning-based aesthetic evaluation of detached housing designs using rendered images

Murat As[1], Imdat As[2]

[1]*Department of Occupational Health and Safety, Aksaray Üniversitesi, Aksaray, Türkiye*
[2]*Department of Architecture, İstanbul Teknik Üniversitesi, İstanbul, Türkiye*

ARTICLE INFO

Article history
Received: 8 January 2025
Revised: 26 August 2025
Accepted: 16 January 2026

Key words:
Automated scoring; aesthetic assessment; deep learning; human scores.

ABSTRACT

The aesthetic evaluation of architectural computer renderings has traditionally remained subjective and dependent on personal, situational, and cultural factors. Within this research, we investigate if deep learning (DL) can be utilized to provide a scientific data-driven solution for approximating the perceived aesthetics in architecture. Our focus is on standalone house designs and uses a dataset of 1,438 computer-rendered competition entries off the Arcbazar website, assigned a rating by professional architects for visual quality. In this research, "aesthetic evaluation" refers to the numerical scores given to the attractiveness to architectural renderings. Our dataset of renderings was standardized through image preprocessing and paired with averaged expert scores. A supervised convolutional neural network (CNN) regression model was then trained and validated using three-fold cross-validation. Model accuracy was established using standard measures of regression (MAE, MSE, RMSE, and R^2). Results indicate that the model was able to predict aesthetic scores with high validity.

While the findings demonstrate the validity of DL models to evaluate architectural renderings, the following limitations should be pointed out: The dependence on rendered views, assessment of just one building type, and expertise based on raters from one platform. Future research will have to expand on the aspects of differing building types, cultural contexts, and multimodal inputs. The incorporation of explainable AI methods will further assist in identifying which visual features contribute most to aesthetic prediction. This work establishes a proof-of-concept framework for integrating deep learning into architectural evaluation, supporting an extensible system that allows for design competition and decision-making. Apart from predictive scoring, such models are well-suited to be integrated with generative design frameworks that will enable the generation of novel architectural proposals optimized in aesthetic quality.

Cite this article as: As, M., & As, I. (2026). Deep learning-based aesthetic evaluation of detached housing designs using rendered images. Megaron, 21(1):32–41.

***Corresponding author**
*E-mail adress: ias@itu.edu.tr

INTRODUCTION

The architectural profession has a long history in developing theories on how to judge built works, for example, in classical antiquity, Vitruvius's *De Architectura* laid out one of the earliest known frameworks for architectural evaluations, emphasizing the principles of *firmitas* (structural stability), *utilitas* (functionality), and *venustas* (aesthetic). These qualities became the building blocks of architectural theory and practice and formed the basis of architectural evaluations for many centuries. In the Renaissance, architects like Palladio defined aesthetic in geometric harmony and proportion. Assessing architectural work has been traditionally done through juried competitions, critical essays, and public opinion surveys. While such approaches are not without value, they are all too often clouded by subjective human judgment and cultural and temporal biases, i.e., evaluating aesthetics is traditionally subjective; we propose a computational approach for detached housing renders.

Recent technological developments may help to overcome some of the limitations of these earlier methods and make it possible for more systematic data-driven approaches to evaluate design. Indeed, evaluating aesthetics in architecture, i.e., visual quality and perceived aesthetic of architectural form, including composition, proportion, and harmony, constitutes one of the most important but inherently subjective issues in design. As artificial intelligence (AI) slowly penetrates modern architectural practice, the need to understand how machine learning (ML) and deep learning (DL) models can assist in quantifying and predicting human-perceived aesthetics has grown. Some research even goes as far as to relate the development of AI in general, to the contribution of aesthetic choices and criteria (Pirozelli & Cortese, 2022). In this paper, we explore the use of DL to predict aesthetic scores for architectural designs with a focus on standalone residential buildings using visual features. We gathered 1,438 images of residential architecture, from Arcbazar.com, an online competition platform, where each building was rated according to aesthetic appeal by regular users, experts, and project-owners.

Our aim is to connect visual architectural features with aesthetic scores with the help of DL models, allowing for a systematic process of design evaluations, i.e. developing a pipeline of preprocessing, convolutional neural networks (CNN) regression modeling, and evaluation metrics. For this purpose, we pre-processed images and standardized them, i.e., create uniform dimensions - and attach human aesthetic scores for a consistent target variable. We developed a DL model and tested it against standard regression metrics, focusing on its potential to predict human scores. The paper is structured as follows: The next section reviews prior work, the methods section describes our methodology, the case study section presents the results of our applied work, the discussion section illustrates the implications, and the conclusion section summarizes our work, discusses limitations and opportunities for future work.

This paper is an early systematic attempt to predict aesthetic scores of detached housing designs using CNN regression, with the ultimate goal to automate the evaluations of aesthetic in a particular design. This opens up new ways in integrating AI into the architectural design workflow and provides an approach for both architects and project-owners to better understand and quantify the visual impact of designs. It has a great potential application in design competitions, e.g., scaling up juries, reducing bias, and supporting smaller firms. We present a DL model, *a baseline framework to bridge subjective human judgments with systematic computational evaluation,* that is trained on human aesthetic scores and offers a robust and scalable framework for evaluating and predicting architectural aesthetic.

BACKGROUND

The use of computational methods in architectural aesthetics has become an important area of research, enabled by the rapidly improving capabilities of AI. Traditionally, aesthetic evaluation is deeply rooted in the cultural, historic, and personal contexts in which it takes place. Nevertheless, recent studies try to quantify aesthetic appeal through computational means with the goal of helping architects, designers, and project-owners reach reliable tools to gauge and foresee the quality of a design. In this section we discuss the theoretical and empirical grounding for AI-powered frameworks for design evaluations.

Aesthetic evaluations have been explored in the domains of psychology, design theory, and architectural practice. Early frameworks, such as those proposed by Vitruvius focused on the triad of structure, function, and aesthetic- aesthetics being a core pillar (Rowland & Howe, 2001). Contemporary theories have extended these ideas to include visual complexity, proportion, symmetry, harmony and contrasting counterparts as critical factors that influence perceived aesthetic (Lorand, 1994; Nasar, 1994). Empirical approaches have also been considered to study human perception of aesthetics (Radwan & Ergan, 2017). For instance, the seminal work of Berlyne (1971) on aesthetic preference implicated arousal potential in the explanation, where humans enjoy stimuli that balance complexity and familiarity. In architecture, such principles are implemented in various design features of a building, from façade composition to materiality and color.

DL models have recently been able to quantify subjective preferences with computational models to extract patterns from visual data (Zhang, 2021; Deng et al, 2017; Elgammal et al, 2017). Traditional methods, which include geometric and rule-based models, have been applied in the assessment of proportions and symmetry within architectural designs

(Stamps, 1999). Such approaches, while successful in respect to particular features, fall short of capturing the holistic nature of aesthetics. DL models have transformed this area of research because they can learn from 'big data' without human intervention, e.g., convolutional neural networks (CNNs) have proved to be very effective in a wide variety of image analysis tasks, ranging from object recognition and style transfer to now aesthetic prediction. Several authors have already used CNNs to assess the visual aesthetics of photography (Lu et al., 2014; Murray, 2012), paintings (Li & Chen, 2009), and art (Saleh & Elgammal, 2015). The application of DL to architectural aesthetics remains largely unexplored. In design, there have been a number of pioneering studies, e.g., an early version involved ML to predict the aesthetics of urban scenes based on greenery, openness of the environment, and building texture (Seresinhe et al., 2017), while other works aimed at extracting visual features from building facades to predict human preferences by means of a supervised learning approach.

Recent advances in DL have enabled more involved analyses. For example, Park et al. (2024) use a DL-based framework to assess wall designs by matching visual features with human aesthetic judgments. Their work has demonstrated just how DL models can learn to represent complex visual patterns which determine aesthetic experience. Other work has explored using GANs in conjunction with generative models to generate designs that are optimized for aesthetic appeal and meaningful designs (Huang et al., 2021). The use of crowdsourced data for aesthetic evaluation has also gained prominence. Platforms like (hidden) have provided valuable datasets for analyzing human preferences. By using ratings from diverse participants these datasets offer robust benchmarks for training and validating predictive models.

However, challenges such as subjectivity, cultural biases, and data inconsistency remain important challenges in developing consistent evaluation tools. Despite notable progress, several gaps persist in the application of AI to architectural aesthetics. First, most of the existing research has focused on specific elements, such as facades or urban layouts, without dealing with the holistic evaluation of stand-alone buildings. Westerdahl et al. (2006), compared the evaluation of 3D virtual reality model to the experience of built edifice, and demonstrated the limitations. Second, even though DL models have an excellent ability in visual feature extraction, the interpretability remains problematic for them, hindering their application in practical workflows of architectural design. Finally, the human factor that may relate to cultural context and emotional response has not received its deserved representation within computational models.

In this paper, we use images of architectural built work, sourced from an online competition platform, to establish a robust framework that links visual features with human-perceived aesthetics. We processed and standardized the images and combined them with their 'aesthetic scores' to ensure the reliability and generalizability of the findings. We also used cross-validation to reduce overfitting and improve the performance of models on diverse datasets. Our work contributes to the fast-emerging domain of computational aesthetics, by introducing a data-driven approach to aesthetics evaluations in architecture. Using AI as a guide in the design process can help architects and project-owners get insight into the visual impact of the design and encourage innovation and creativity in the profession.

METHODS

In this study, we followed a supervised learning approach, which included: Clearly defined data collection steps, image preprocessing, integration of expert-labeled scores, training a deep learning (DL) regression model, and model evaluation, as shown in Figure 1. We used a dataset of 1,438 residential building images from arcbazar.com, where expert architects rated each building's visual aesthetic appeal, resulting in a numeric aesthetic score associated with each image. Although computer renderings provide controlled visual consistency they cannot fully replicate material, at-

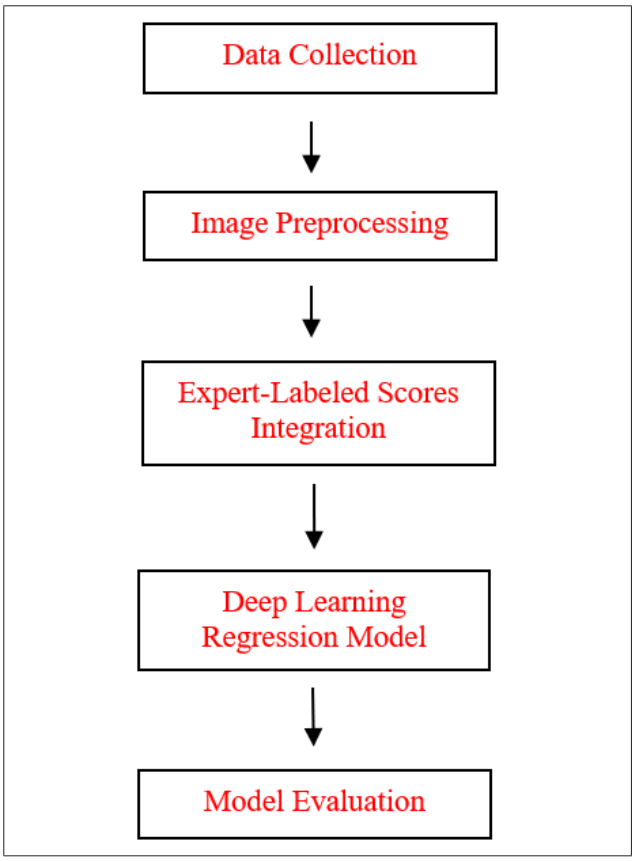

Figure 1. Deep learning-based workflow for aesthetic evaluation of architectural images.

mospheric, or spatial qualities present in photographs of built works. However, they provide standardized views across competitions, they avoid issues with varied photo quality, e.g., lighting, angle, environment, which allowed us to work with a consistent dataset, and also they reflect common practice in architectural competitions, at which state such aesthetic evaluations become significant for stakeholders to make decisions.

We studied the regression relationship between visual features of the buildings and their corresponding beauty scores given by experts. All images were resized to 224×224 pixels and converted to PNG format to ensure consistent dimensions and channel structure for deep learning input. That was a vital step since all images needed to be consistently matched in resolution and color channels for easy training. The target variable was the expert-assigned aesthetic score, averaged per image, used as the regression output for the model.

Image preprocessing steps, such as resizing and format conversion, were performed using Jupyter Notebook scripts. Then, we tested the DL model for predicting the individual building's aesthetic score against its image using a deep neural network in MATLAB.

We used 3-fold cross-validation for performance evaluation to ensure each sample was used in both training and testing phases. This choice was made to balance between computational efficiency and variance estimation, as 3-fold CV offers a good trade-off for relatively moderate-sized datasets. We fine-tuned the model by utilizing initial learning rate limits, i.e. a small number that controls how big a step the model takes when adjusting itself to improve accuracy --too high = unstable, too low = very slow learning, maximum number of epochs, where an epoch means one complete pass of the training data through the model, and adjusting the batch size. We evaluated how well the predicted aesthetic scores matched the ground-truth expert scores using statistical accuracy metrics. We measured model accuracy using MAE, MSE, RMSE, and R^2. MAE gives the average absolute prediction error; MSE penalizes larger errors; RMSE gives an interpretable magnitude of error; and R^2 explains the variance captured by the model. These numbers measure how close the model's predictions are to the real values. Lower error values mean better predictions, and R^2 close to 1 means the model explains most of the variation. In simple terms, these values measure how far the model's predictions are from actual expert scores, with lower values meaning higher accuracy.

To ensure stability and reliability, model performance was evaluated using 3-fold cross-validation, with accuracy variation across folds used to assess generalization. This means the dataset was split into three parts and the model was trained and tested three times so that every sample was used for both purposes. This reduces the risk of overfitting and ensures fair testing. Our deep learning code first constructs input data by linking image paths with a target variable representing the beauty score. The dataset is then randomly shuffled to ensure a balanced distribution between training and testing sets.

The code in Figure 2 splits the data into three parts, with one part being used for testing and the remaining two used for training in each iteration using the cvpartition function. Various performance metrics-MSE, RMSE, MAE, and R^2-are declared as empty arrays. Training and testing for each fold require the proper distribution of datasets in every iteration using the training and test functions. Training the DL model requires using the trainNetwork function. Hyperparameters include learning rate, epoch count, and mini-batch size. These are defined by the trainingOptions function. The model architecture lgraph_1 is predefined as a layerGraph object. Images from both the training and testing datasets are read and converted into 4-dimensional tensors to match the input of the network. Here, the model predicts the images of training and testing datasets using the predict function. For computing the metrics, a custom function called calculate_metrics2 is used to measure the errors between the actual values and predictions made by the model. For each fold, MSE, RMSE, MAE, and R^2 are

```
% K-fold cross-validation parameters
K = 3;
c = cvpartition(size(data, 1), 'KFold', K);

% Create empty arrays for performance metrics
trainMSEs = zeros(K, 1);
trainRMSEs = zeros(K, 1);
trainMAEs = zeros(K, 1);
trainR2s = zeros(K, 1);
testMSEs = zeros(K, 1);
testRMSEs = zeros(K, 1);
testMAEs = zeros(K, 1);
testR2s = zeros(K, 1);

for k = 1:K
    % Split the data into training and testing sets
    train_data = data(training(c, k), :);
    test_data = data(test(c, k), :);

    % Set up training options for the neural network
    options = trainingOptions("adam", ...
        "InitialLearnRate", initialLearnRate, ...
        "MaxEpochs", maxEpochs, ...
        "MiniBatchSize", miniBatchSize, ...
        "Shuffle", "every-epoch", ...
        "Verbose", false);

    % Train the network
    ccnet = trainNetwork(train_data, lgraph_1, options);
```

Figure 2. Code developed for 3-fold cross validating an image-based regression DL model.

computed and stored in corresponding arrays. The final step is reporting the results by calculating all the fold's mean and standard deviation for both the training and testing datasets.

MODELING AESTHETIC EVALUATIONS WITH DEEP LEARNING

This case study is based on design entries submitted to residential architectural competitions hosted on arcbazar.com. Our goal was to develop a DL model that can automatically evaluate architectural design work. We explored the predictive performance of our model using various statistical metrics, e.g., Mean Absolute Error, and R-squared. Model robustness was validated through 3-fold cross-validation, residual analysis, and comparison of predicted vs. actual expert scores.

Figure 3 illustrates the full workflow, from data collection and image preprocessing to model training and final performance evaluation. We trained our DL model with 1438 images from residential competition results at (hidden). Each image had beauty scores from three different types of users, e.g., regular users, expert architects who already won a certain number of competitions, and project-owners who actually launch the competitions. For our study we only used the scores of expert architects, since scores of regular users showed too much variation. The predicted scores closely approximated the actual scores given by expert architects, indicating strong regression performance.

As shown in Table 1, the performance metrics of the model are presented, including MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), and R^2 (R-squared), which are evaluated using 3-fold cross-validation. MAE measures the average magnitude of errors between predicted and actual values, providing an intuitive indication of model performance. MSE calculates the average of the squared differences between predictions and observations, penalizing larger errors more heavily and making it sensitive to outliers. RMSE, as the square root of MSE, expresses the error in the same units as the target variable, offering a practical interpretation of predictive accuracy. R^2, also known as the coefficient of determination, quantifies the proportion of variance in the dependent variable explained by the independent variables, with values closer to 1 indicating a better fit. Cross-validation, a robust resampling technique, is used to assess model performance by splitting the dataset into multiple folds and alternating between training and validation sets, ensuring that every data point is utilized in both phases. Additionally, the model was trained using an initial learn rate of 0.0005, which controls the step size of weight updates during optimization, with a maximum of 10 epochs, denoting the total number of complete passes through the dataset, and a mini-batch size of 64, specifying the number of samples used per iteration. These training parameters and metrics collectively provide a comprehensive assessment of the model's performance and training configuration.

The metrics used for the training data reflect the model's performance. The MAE, MSE, and RMSE values are relatively low: MAE is 0.3324, MSE is 0.199, and RMSE is 0.4282, indicating high accuracy of the model on the training data. The R-squared (R^2) value is 0.9745, showing that the model explains over 97.45% of the variance in the target variable, demonstrating excellent performance on the training data.

We used the same metrics (MAE, MSE, RMSE, and R^2) to evaluate the model's performance on unseen test data. The errors on the test data are higher than those on the training data, with MAE being 0.7618, MSE 1.2947, and RMSE 1.1377. However, all these values remain within acceptable limits and reflect reasonable generalization capability. The average R^2 value of 0.8340 suggests that the model describes about 83.40% of the variance in the target variable on the test data.

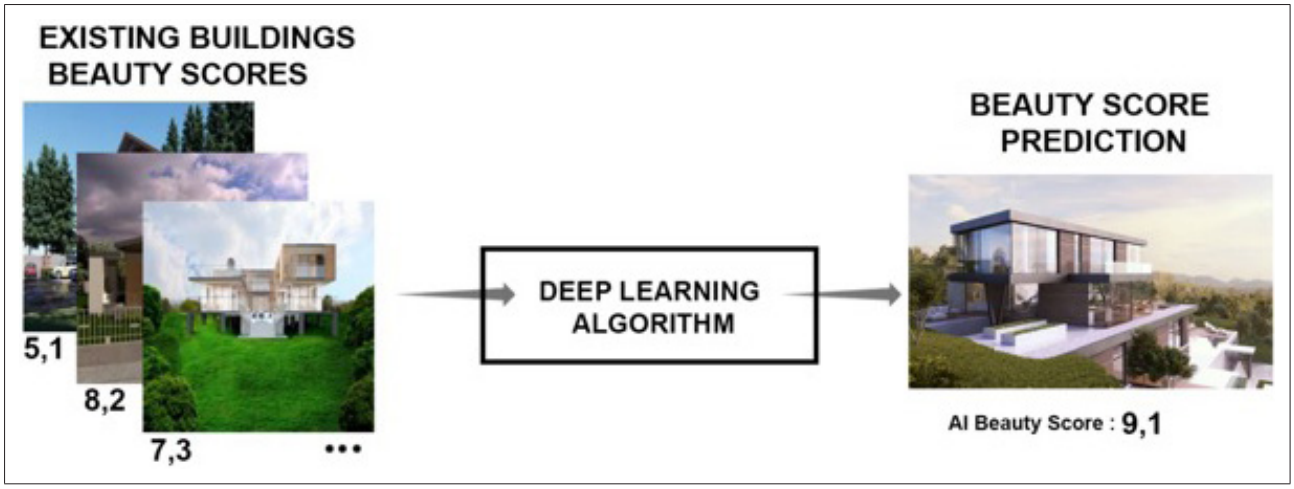

Figure 3. Basic DL workflow for predicting the aesthetic scores of images.

Table 1. Performance Metrics and Hyperparameters of Training and Testing Data Based on Cross-Validation Results

Performance Metrics	Train				Test			
	Mae	Mse	Rmse	R^2	Mae	Mse	Rmse	R^2
Cross_Validation_1	0.2649	0.1140	0.3376	0.9856	0.7383	1.2870	1.1345	0.8293
Cross_Validation_2	0.4651	0.3659	0.6049	0.9534	0.7913	1.2506	1.1183	0.8380
Cross_Validation_3	0.2672	0.1171	0.3422	0.9847	0.7558	1.3465	1.1604	0.8346
Average	0.3324	0.199	0.4282	0.9745	0.7618	1.2947	1.1377	0.8340
Initial Learn Rate	0.0005							
Max Epochs	10							
Mini Batch Size	64							

The cross-validation results also depict a small variance in performance between the best and worst folds, especially with regard to R^2 and MAE. For example, the R^2 values of the test data vary between 0.8293 and 0.8380, meaning that the model performed very well across the different folds. Hence, the model both on the training and testing side performs well.

Considering the R-squared is high-as close to 83.40%- we can conclude that from the test dataset, the actual value of a target variable can be well predicted through the model itself.

The blue circles in the top-left plot of Figure 4 represent the relationship between the observed and predicted values in

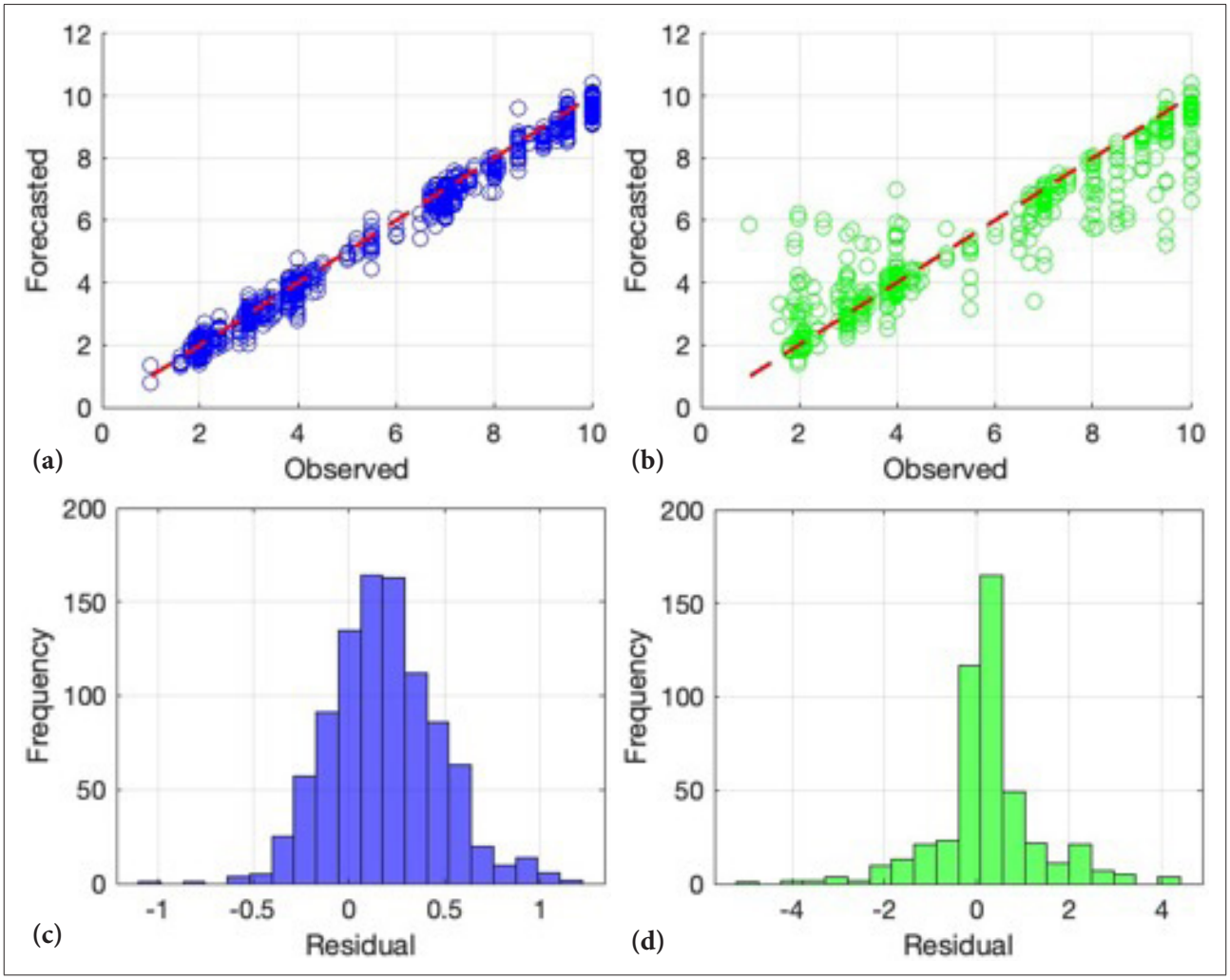

Figure 4. Evaluation of Model Performance: Relationships between predicted and observed values for training **(a)** and test data **(b)**, and histograms of residual distributions for training **(c)** and test data **(d)** are presented.

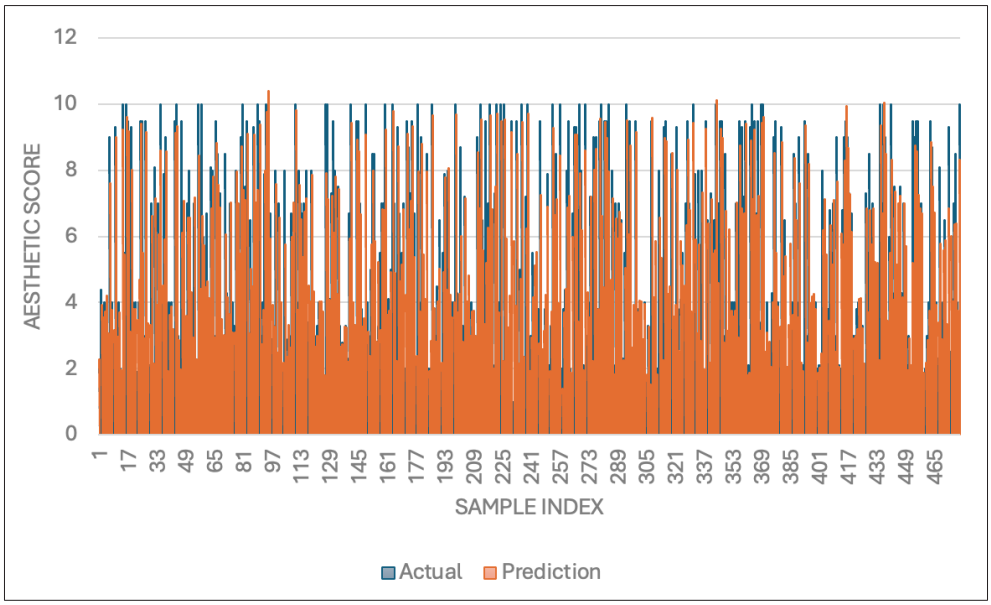

Figure 5. Comparison of actual and predicted values obtained from the test dataset.

the model's training data. The red dashed line indicates the ideal prediction line, y=x, perfect prediction, which allows us to visually assess how accurate the model's predictions are. The predicted values are close to the observed values, which means that the model performs well with the training data. In the top right plot, the green circles show the relation between the predicted and actual values in the test data. Again, the red dashed line represents the ideal line of prediction. The predicted values are generally close to the ones observed, although there is a bit of a more spread in these points compared to the training data. That suggests the performance of the model is somewhat worse when it comes to this test data.

The bottom-left histogram shows the frequency distribution of the model's prediction errors (deviations) on the training data. There is a crowding of the deviations around 0, which means that for the most part, the model predicts without large errors. The shape of the histogram resembles a normal distribution, which suggests that the errors are random and there is no systematic bias in the model. Similarly, the bottom-right histogram presents the distribution of prediction errors on the test data. The deviations are also concentrated around 0 but with a wider spread than for the training data. The errors in the test data are more variable, indicating that the model struggled more when it had to predict new images.

In general, the model shows great performance with the training data. The values are well matched between predicted and observed, and there is very little error. For the test data, as shown in Figure 5, the model performs well with slightly more errors and dispersion – compared to the training data. The graph shows that the actual values, Test_actual, and the predicted values, Test_prediction, are normally on a very similar path and clearly indicate that the model performed well in its predictions. The overall trends within the dataset have been well captured by the model, as reflected in its capability to generate predictions that are

Table 2. Actual and predicted values for the first 20 data points out of a total of 479 observations in the test dataset

Data	Actual	Prediction
1	2.1	2.2773921
2	4.4	3.9213858
3	3.8	3.5455017
4	4	3.7227278
5	3	4.2232203
6	1.8	3.0767283
7	9	7.617538
8	4	3.7764783
9	3	3.174438
10	9.3	9.0082617
11	3.3	2.961113
12	4	3.7118716
13	2	1.961212
14	10	9.245945
15	5.5	5.4527626
16	10	9.619029
17	9.5	9.1773605
18	2.4	3.121088
19	9.3	8.0058241
20	4	3.7962461

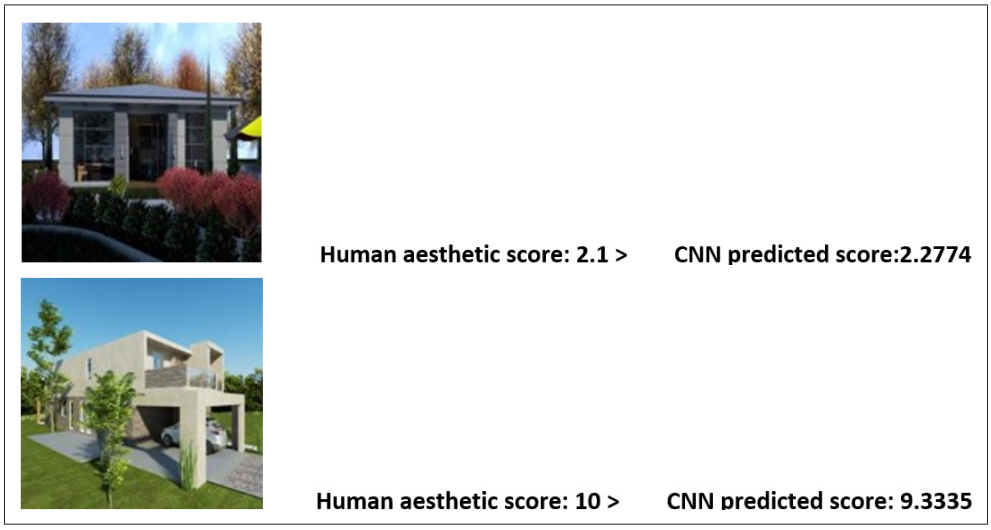

Figure 6. Shows two examples of the home designs scored on (hidden). You see the actual image and its human score, and to its right you see the aesthetic score predicted from our DL model.

largely consistent with the actual values. Notably, the deviations of the model outputs from the actual values are mostly minimal, indicating how reliable the model is.

Overall, as seen in the first image of Figure 6, where the human aesthetic score is 2.1, our deep learning model predicted 2.2774, and for the second image, where the human aesthetic score is 10, it predicted 9.3335. The model shows a great level of accuracy in its predictive performance. Most of the predicted values lie very close to the actual values, which means that the model has grasped the underlying patterns in the data.

These results underpin the high predictive capability of the model and also extend its applicability to more complex datasets or diverse scenarios where its robust performance can be tapped into effectively evaluating built work.

As seen in Table 2, the comparison between actual values and model predictions for the images is presented. The "Data" column represents the image identifier, while the "Actual" column shows the true values or observations, and the "Prediction" column displays the values predicted by the model. For example, for the first image (Data 1), the actual value is 2.1, and the model predicts a value of 2.2773921. Similarly, for Image 2, the actual value is 4.4, while the prediction is 3.9213858, and this continues for the first 20 images. This table demonstrates the closeness of the predicted values to the actual ones for the first twenty test data points.

DISCUSSION

This study demonstrates the potential of convolutional neural networks (CNNs) to reliably predict human-rated aesthetic scores of detached house designs. The results confirm the capability of supervised deep learning algorithms to learn correlations between rendered architectural images and professional judgments. These findings show the potential of computational techniques to enhance traditional architectural evaluations, particularly in settings such as design competitions where large numbers of submissions need to be evaluated.

Human Raters and Evaluation Framework
Professional architects in the Arcbazar competition platform offered aesthetic ground-truth values utilized in the research. These members had verifiable professional background and prior recognition in the site, they collected a certain number of points by winning architectural competitions ran on the platform. All the designs were given numerical scores of visual attractiveness on a continuous scale. While these scores were a good proxy for "aesthetic value," they remain subjective and dependent upon the raters' experience and cultural perspectives. The model was trained on the aggregate of individual opinions rather than objective principles. Future research could benefit from the application of more formal assessment rubrics or several groups of raters to reduce bias.
Complex Patterns and Model Learning

The results demonstrate that CNNs can learn complex architectural features related to perceived aesthetic, which encompass façade composition, proportional relations, rhythm, balance between solid and void, and overall harmony of massing. While such visual characteristics adhere to dominant theories of architectural aesthetic (e.g., Stamps, 1999; Lorand, 1994), the model's outputs were statistically derived –and, not via design rules, codes etc. This demonstrates the power of deep learning to detect implicit aesthetic characteristics embedded in architectural imagery.
Interpretability and Explainable AI
The weakness of this study is interpretability of predictions. Capable though they are, CNNs provide little insight into

which parts of an image or features drive their predictions, researchers work on interpreting and understanding DL models (Samek, 2017). Future work should utilize explainable AI methods, e.g., Gradient-weighted Class Activation Mapping (Grad-CAM) and Shapley Additive Explanations (SHAP) visualizations—to demarcate model attention regions and allow predictions to align with architectural reasoning (Zheng, 2022). Such tools would enable it to determine, for example, if the model gives more importance to façade symmetry, roof ratio, or window arrangement when it generates aesthetic score increment predictions.

Disadvantages of Render-Only Approach

This research is limited in its reliance on rendered representations of individual detached homes. Renders offer standardization and visual simplicity, and they are the main media type when it comes to stakeholders making decision whether to take a design to the next level, i.e., to build it or not. However, they may not convey atmospheric richness, spatiality, or cultural specificity—all factors demonstrated to have an impact on aesthetical perception (Bille, M., & Sorensen, 2016; Herzog & Shier, 2000; Radwan & Ergan, 2017). The data set was also sampled from a single platform and dwelling type. Generalizability of this model demands increased data sets that cover images of built works, mixed building types, and cross-cultural testing.

Implications and Future Directions

Despite these constraints, the work presents a venue for computational aesthetic evaluation in architecture. Directions for future work are:

- Expansion of data sets involving diverse building types and cultural contexts.

- Incorporation of multimodal data (e.g., text description, environmental context, VR/AR spatial experience).

- Incorporation of explainable AI techniques to improve transparency.

- Synthesis of predictive models and generative techniques (e.g., GANs, diffusion models) to enable the creation of new designs optimized for aesthetic value.

With these strategies, AI systems can potentially go beyond acting as predictive critics to become engaged design collaborators, and offer architects and clients new tools for understanding, contrasting, and generating aesthetically informed designs.

CONCLUSION

This paper described a deep learning–oriented methodology for predictive aesthetic judgments of detached house designs from rendered images as input. By training a convolutional neural network on 1,438 human-rated renders, derived from competition submissions on Arcbazar, we demonstrated the feasibility of computational approximation to human-perceived aesthetic with high predictive accuracy. The results confirm the feasibility of subjective judgments' quantification using supervised learning and indicate the potential of artificial intelligence for complementing traditional design evaluation methods.

The study makes three central contributions. First, it lays down a reproducible pipeline to align architectural imagery with aesthetic scores by pre-processing, supervised CNN training, and systematic validation. Second, it suggests detached housing renders as a controlled test case and thereby lays the ground for applying the research to other building types and larger datasets. Third, it identifies opportunities and challenges of computational aesthetics—namely, the need for interpretability, cross-cultural validation, and multimodal design data integration.

There are several limitations to take into account. Render-only images, while standardized, will be unable to capture the full depth of architectural aesthetics that include materiality, spatial experience, and cultural meaning. The dataset's reliance on one platform and set of experts further restricts generalizability. Such limitations indicate that the results need to be taken as proof-of-concept and not a definitive model of architectural aesthetics.

In the future, promising paths include applying explainable AI methods to reveal which visual properties have the most significant impact on predictions, multimodal input such as VR/AR or text descriptions to enhance representational richness, and generative modeling (GANs, diffusion models) to create aesthetically polished design proposals. As a whole, these advancements have the potential to transform AI systems from evaluation tools into generative co-creatives—enhancing human judgment, making it more transparent, and enabling more informed and participatory architectural decision-making.

ETHICS: There are no ethical issues with the publication of this manuscript.

PEER-REVIEW: Externally peer-reviewed.

CONFLICT OF INTEREST: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

FINANCIAL DISCLOSURE: The authors declared that this study has received no financial support.

REFERENCES

Berlyne, D. E. (1971). *Aesthetics and psychobiology*. Appleton-Century-Crofts.

Bille, M., & Sørensen, T. F. (2016). *Elements of architecture: Assembling archaeology, atmosphere and the perfor-*

mance of building spaces (1st ed.). Routledge. https://doi.org/10.4324/9781315641171

Deng, Y., Loy, C. C., & Tang, X. (2017). Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine*, *34*(4), 80–106. https://doi.org/10.1109/MSP.2017.2696576

Elgammal, A., Liu, B., Elhoseiny, M., & Mazzone, M. (2017). CAN: Creative adversarial networks, generating "art" by learning about styles and deviating from style norms. *arXiv, 2017,* 07068.

Herzog, T. R., & Shier, R. L. (2000). Complexity, age, and building preference. *Peer Reviewed Articles*, 48, 557–575. https://doi.org/10.1177/00139160021972667

Huang, J., Johanes, M., Kim, F. C., Doumpioti, C., & Holz, G. C. (2021). On GANs, NLP and architecture: Combining human and machine intelligences for the generation and evaluation of meaningful designs. *Technology Architecture and Design*, *5*(2), 207–224. https://doi.org/10.1080/24751448.2021.1967060

Li, C., & Chen, T. (2009). Aesthetic visual quality assessment of paintings. *IEEE Journal of Selected Topics in Signal Processing*, *3*(2), 236–252. https://doi.org/10.1109/JSTSP.2009.2015077

Lorand, R. (1994). Beauty and its opposites. *The Journal of Aesthetics and Art Criticism*, *52*(4), 399–406. https://doi.org/10.1111/1540_6245.jaac52.4.0399

Lu, X., Lin, Z., Jin, H., & Yang, J. (2014). RAPID: Rating pictorial aesthetics using deep learning. In *Proceedings of the 22nd ACM International Conference on Multimedia, 2014,* 457–466. https://doi.org/10.1145/2647868.2654927

Nasar, J. L. (1994). Urban design aesthetics: The evaluative qualities of building exteriors. *Environment and Behavior*, *26*(3), 377–401. https://doi.org/10.1177/001391659402600305

Murray, N., Marchesotti, L., & Perronnin, F. (2012). AVA: A large-scale database for aesthetic visual analysis. *2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012,* 408–415. https://doi.org/10.1109/CVPR.2012.6247954

Park, S. B., Park, J. H., & Jung, S. (2024). Comparative aesthetic assessment of machine learning and human judgment for building wall designs. *Architectural Science Review*, *67*(4), 321–331. https://doi.org/10.1080/00038628.2023.2278500

Pirozelli, P., & Cortese, J. F. (2022). The beauty everywhere: How aesthetic criteria contribute to the development of AI. In *I (Still) Can't Believe It's Not Better! Workshop at NeurIPS 2021* (pp. 69–74). PMLR.

Radwan, A., & Ergan, S. (2017). Quantifying human experience in interior architectural spaces. *ASCE International Workshop on Computing in Civil Engineering, 2017*, 373–380. https://doi.org/10.1061/9780784480830.046

Rowland, I. D., & Howe, T. N. (2001). *Vitruvius: Ten books on architecture*. Cambridge University Press.

Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv, 2017,* 08296.

Saleh, B., & Elgammal, A. (2015). Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv, 2015,* 00855.

Seresinhe, C. I., Preis, T., & Moat, H. S. (2017). Quantifying the impact of scenic environments on health. *Scientific Reports*, *5*, 16899. https://doi.org/10.1038/srep16899

Stamps, A. E. (1999). Physical determinants of preferences for residential facades. *Environment and Behavior*, *31*(6), 723–751. https://doi.org/10.1177/00139169921972326

Westerdahl, B., Suneson, K., Wernemyr, C., Roupé, M., Johansson, M., & Allwood, C. M. (2006). Users' evaluation of a virtual reality architectural model compared with the experience of the completed building. *Automation in Construction*, *15*(2), 150–165. https://doi.org/10.1016/j.autcon.2005.02.010

Zhang, J., Miao, Y., & Yu, J. (2021). A comprehensive survey on computational aesthetic evaluation of visual art images: Metrics and challenges. *IEEE Access*, *9*, 77164–77187. https://doi.org/10.1109/ACCESS.2021.3083075

Zheng, Q., Wang, Z., Zhou, J., & Lu, J. (2022). Shap-CAM: Visual explanations for convolutional neural networks based on Shapley value. In *European Conference on Computer Vision* (pp. 459–474). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-19775-8_27