Residual Fusion of Tabular Data and Satellite Imagery for Property Price Estimation

Prabhat Chandra Tiwari (24114066)

1 Overview

Property prices depend on more than just the features of the house itself, such as size, number of rooms, or construction quality. They are also strongly influenced by the surrounding neighborhood, including road access, nearby buildings, green spaces, and overall development of the area. While tabular data captures the internal characteristics of a property well, it does not fully represent this neighborhood context.

In this project, we aim to improve property price prediction by combining tabular data with satellite imagery. Satellite images provide a visual view of the neighborhood around each property and naturally capture information such as urban density, connectivity, and land use, which are difficult to describe using numerical features alone.

Our approach uses a two-stage modeling strategy. First, we train an XGBoost model using only tabular features. This model acts as a strong baseline and learns most of the predictable patterns related to property size, location, and amenities. Latitude and longitude are included to capture coarse spatial information.

Next, instead of directly mixing image and tabular features, we use a residual learning approach. A Convolutional Neural Network (CNN) is trained on satellite images to predict a correction term, called a residual. This residual represents how much the XGBoost prediction should be adjusted based on visual neighborhood information.

The final prediction is computed as:

$$\text{Final Prediction} = \text{XGBoost Prediction} + \text{CNN Residual}$$

We also experimented with learning an additional scaling parameter on the residual contribution; however, this did not improve performance and is therefore treated as an ablation rather than the final model. This design is chosen because simple fusion methods, such as directly concatenating image and tabular features, often perform poorly. In our experiments, naive fusion reduced accuracy because the strong tabular signal dominated the learning process, preventing the image model from contributing effectively.

By using residual fusion, each model plays a clear role. The XGBoost model provides a reliable base estimate, while the CNN focuses only on correcting mistakes related to neighborhood context. This results in better performance, more stable training, and easier interpretation of how satellite images influence the final prediction.

2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is conducted to understand the structure of the dataset and the behavior of property prices before modeling. This analysis focuses on examining the distribution of

the target variable, identifying skewness and outliers, and exploring relationships between tabular features and property prices.

By analyzing these patterns, we gain insights into which structured features are most informative and identify limitations of tabular data alone. These observations help motivate the use of additional contextual information in later sections.

2.1 Price Distribution

We begin by analyzing the distribution of property prices. Since raw prices are highly skewed and contain extreme outliers, we apply a logarithmic transformation to stabilize variance and improve model behavior. Figure 1 shows the distribution of log-transformed property prices. To further examine the statistical properties of the transformed target, we use a Q–Q plot to compare the empirical distribution with a normal distribution. As shown in Figure 2, the log-transformed prices follow an approximately normal distribution in the central region, with mild deviations at the tails.

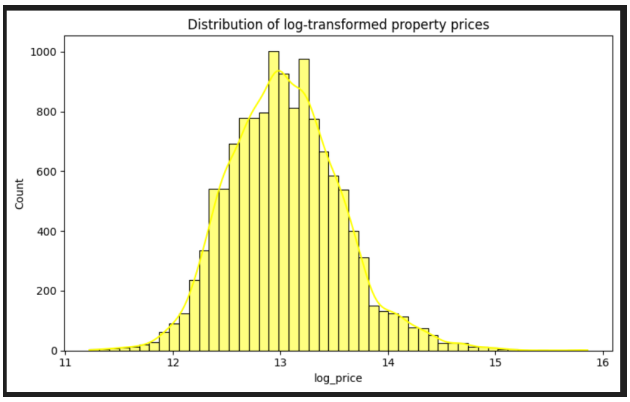

Figure 1: Distribution of log-transformed property prices

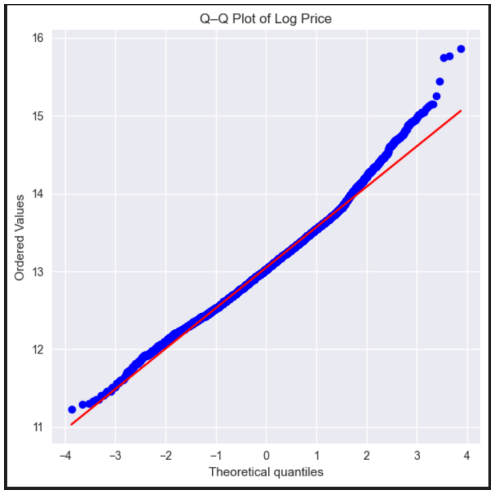

Figure 2: Q–Q plot of log-transformed property prices

2.2 Relationship Between Features and Price

Next, we analyze the relationship between tabular features and property prices. Figure 3 presents the features with the strongest linear correlation to log-transformed price. Size-related attributes such as living area and property grade show the highest correlation, which aligns with domain intuition. Some features are moderately correlated with each other, indicating potential multicollinearity. Tree-based models handle this well, while neural models benefit from learned representations. However, correlation strength alone does not capture non-linear interactions or neighborhood-level effects, motivating the use of more expressive models and additional contextual information.

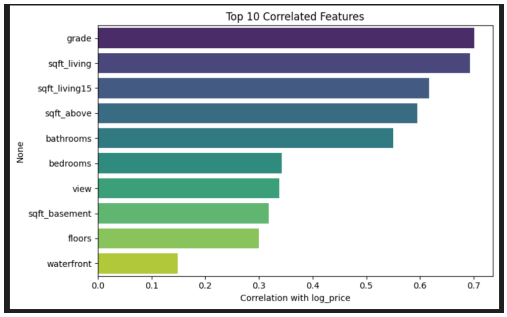

Figure 3: Top tabular features correlated with log-transformed prices

2.3 Spatial Distribution of Property Prices

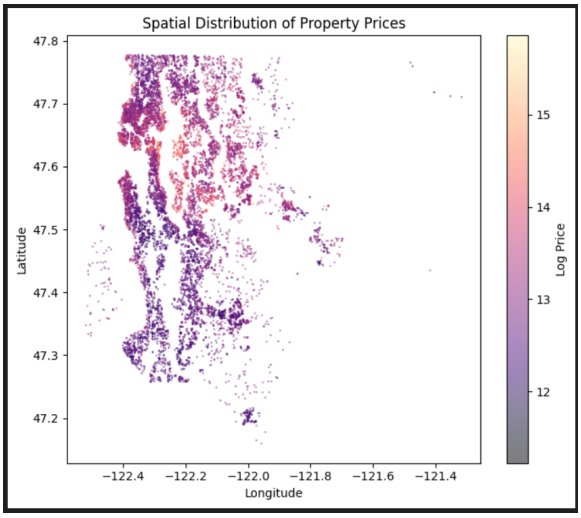

Figure 4: Spatial distribution of log-transformed property prices

This image visualizes the spatial distribution of log-transformed property prices using geographic coordinates. Each point represents a property, positioned by its latitude and longitude, with color indicating the corresponding log price value. The plot reveals clear spatial clustering, where higher-priced properties are concentrated in specific geographic regions, while lower-priced properties are more dispersed across other areas.

These visible price gradients indicate that property values vary systematically with location and neighborhood context rather than being uniformly distributed.

This spatial heterogeneity highlights the importance of geographic and surrounding contextual information in property valuation and motivates the use of satellite imagery to capture neighborhood-level visual cues beyond tabular location features.

3 Financial and Visual Insights

While tabular features capture important structural attributes of a property, they do not fully describe the surrounding neighborhood context that can influence property value. To better understand how satellite imagery contributes to price estimation, we analyze the visual component of the model using interpretability techniques.

In this section, we focus on the residual CNN and examine how visual cues from satellite images influence price corrections beyond the tabular baseline. Using Grad-CAM visualizations, we identify which regions of the images are most influential and analyze how these visual factors relate to upward or downward adjustments in predicted prices.

3.1 Understanding Visual Contributions

Satellite imagery provides a direct visual representation of this surrounding context. To interpret how visual information contributes to price estimation, we apply Grad-CAM to the residual CNN. This technique highlights spatial regions in satellite images that most strongly influence the model's residual corrections, allowing us to analyze which visual factors drive price adjustments.

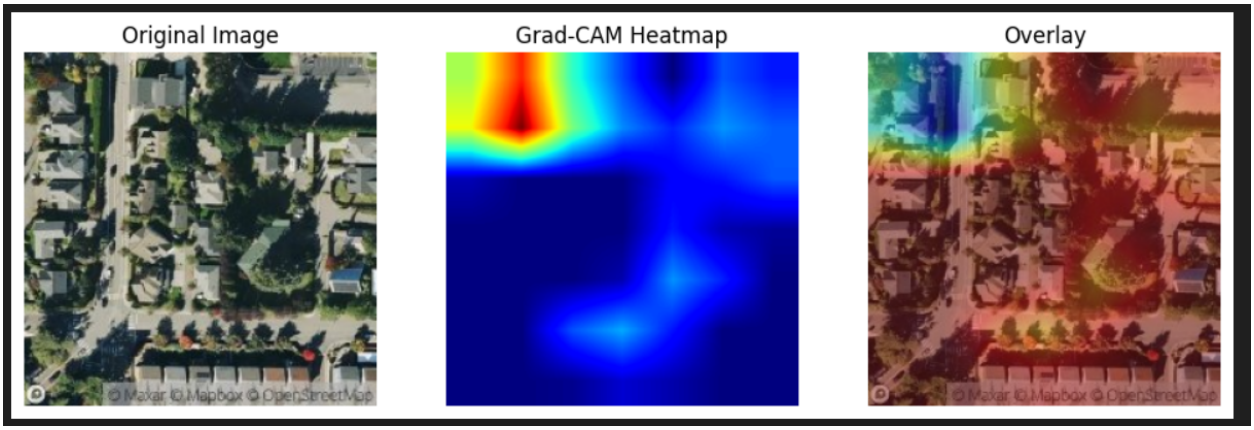

Figure 5: Grad-CAM visualization for a representative validation sample showing original image, attention heatmap, and overlay

We begin by analyzing Grad-CAM visualizations for individual validation samples. This example illustrates how the residual CNN allocates attention across the surrounding neighborhood rather than focusing solely on the property footprint.
The highlighted regions correspond to nearby development patterns and surrounding infrastructure. Grad-CAM visualizations show that the CNN consistently focuses on meaningful spatial features present in satellite images. These include road networks, building density, and the overall layout of the surrounding neighborhood. Such features are directly related to accessibility, connectivity, and urban development, all of which are important drivers of property value. We visualize zoom16 for Grad-CAM as it captures fine-grained local structure

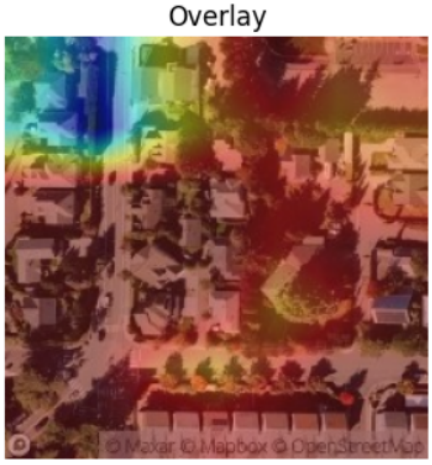

Figure 6: Grad-CAM visualization for a single validation sample

3.2 Positive and Negative Residual Analysis

We next perform a contrastive analysis by comparing Grad-CAM maps for properties where the tabular model significantly underestimates prices (positive residuals) and overestimates prices (negative residuals). This comparison allows us to study the directional influence of visual cues. The visualizations reveal clear differences between the two cases.In several examples, positive residuals are associated with greener and less congested surroundings, while negative residuals frequently emphasize dense built environments. These patterns suggest that the model captures environmental quality signals not present in tabular features. This demonstrates that visual context contributes not only to prediction accuracy but also to the direction of price adjustment.

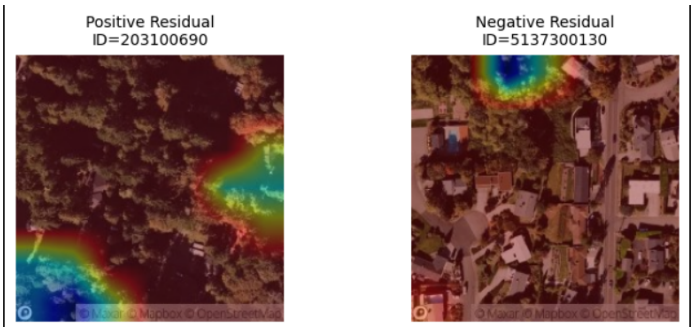

Figure 7: Grad-CAM comparison for positive (left) and negative (right) residual cases

3.3 Dominant Visual Factors in High-Impact Corrections

While some properties are well predicted using tabular features alone, others exhibit large prediction errors where additional information becomes critical. To identify which visual factors influence the model most strongly, we focus on samples where the tabular model produces the largest absolute errors. In these high-impact cases, the image-based correction plays a significant role in the final prediction.

Figure 8 shows the average Grad-CAM visualization computed over these high-impact samples. This map highlights image regions that consistently influence large residual corrections across multiple properties.

The dominant attention patterns emphasize dense built-up regions, road connectivity, and surrounding neighborhood structure. These visual cues appear repeatedly across high-impact cases, indicating that neighborhood-level context is among the most influential visual factors used by the model when correcting tabular predictions.

Unlike individual Grad-CAM examples, this dominant visualization captures stable and consistent visual drivers of price correction, providing stronger evidence of which visual factors influence the model most.

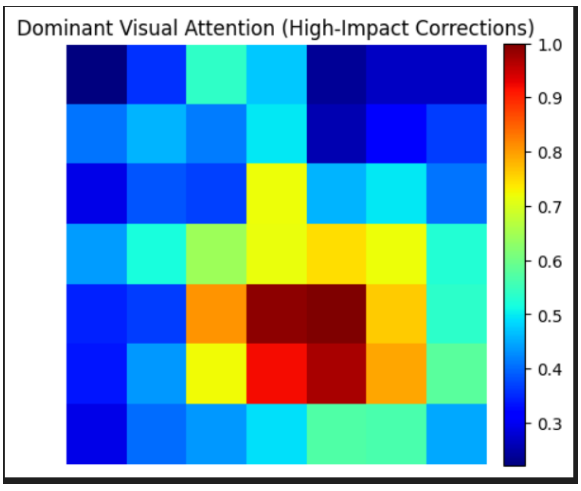

Figure 8: Dominant Grad-CAM attention over high-impact residual corrections

- Warmer colors (yellow to red) indicate image regions that strongly influenced the model's residual correction.
- Cooler colors (blue) represent regions with lower influence on the prediction.
- The color scale reflects relative importance rather than absolute contribution.
- As the visualization is averaged across high-impact samples, it highlights consistent neighborhood-level patterns instead of features from a single property.

3.4 Aggregate Grad-CAM and Robustness of Visual Factors

While the analysis of high-impact cases reveals which visual cues influence large residual corrections, it is also important to verify whether these cues are consistently used across the dataset. To assess robustness, we compute an aggregate Grad-CAM by averaging attention maps across multiple validation samples.

The resulting visualization highlights image regions that, on average, receive attention from the residual CNN across many properties. This provides insight into the general visual patterns learned by the model rather than sample-specific behavior.

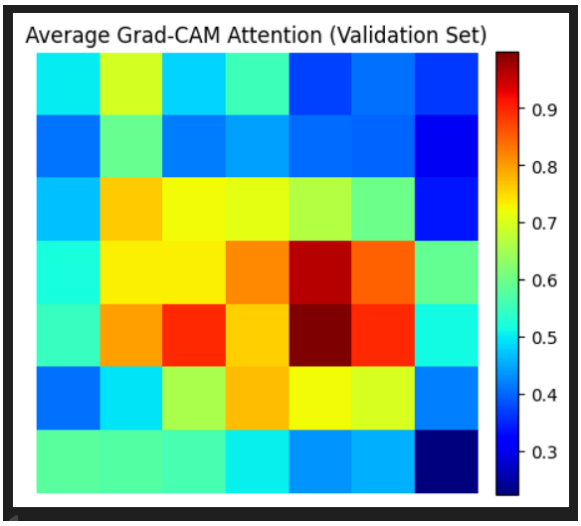

Figure 9: Average Grad-CAM attention across validation samples

- Warmer colors (yellow to red) indicate image regions that consistently influence the model across many validation samples.
- Cooler colors (blue) represent regions with lower average influence on price correction.
- The smooth and structured attention pattern reflects stable neighborhood-level visual cues learned by the model.
- As this visualization is averaged across the validation set, it highlights generalizable visual factors rather than sample-specific details.

4 Model Architecture

Figure 10 illustrates the overall multimodal architecture used in this project. The model consists of two parallel components: a tabular regression branch and an image-based residual correction branch.

The tabular branch takes structured input features such as property size, number of rooms, amenities, and geographic coordinates. These features are passed to an XGBoost regressor, which produces a baseline prediction of the log-transformed property price. This model captures strong non-linear relationships present in structured data and serves as a stable initial estimator.

In parallel, satellite images of the property surroundings are processed by a convolutional neural network (CNN). The CNN acts as an image encoder that extracts high-level visual features related to neighborhood structure and surrounding development. Rather than predicting prices directly, the CNN outputs a residual correction that represents information not captured by the tabular model.

The residual fusion layer combines the baseline tabular prediction with the CNN-predicted residual. This design ensures that the image model focuses specifically on correcting errors made by the tabular model, reducing the risk of overfitting and improving interpretability.

The architecture separates structured and visual information into two branches to leverage their complementary strengths. XGBoost is used for tabular features due to its strong performance on

structured data, while the CNN processes satellite imagery to capture neighborhood-level visual context.

Instead of predicting prices directly from images, the CNN is trained to estimate residual corrections to the tabular model. This residual learning strategy constrains the visual branch to focus on information not already captured by tabular features, leading to more stable and effective multimodal integration.

The final output of the architecture is a multimodal log-price prediction that integrates both structured and visual information in a principled and modular manner.

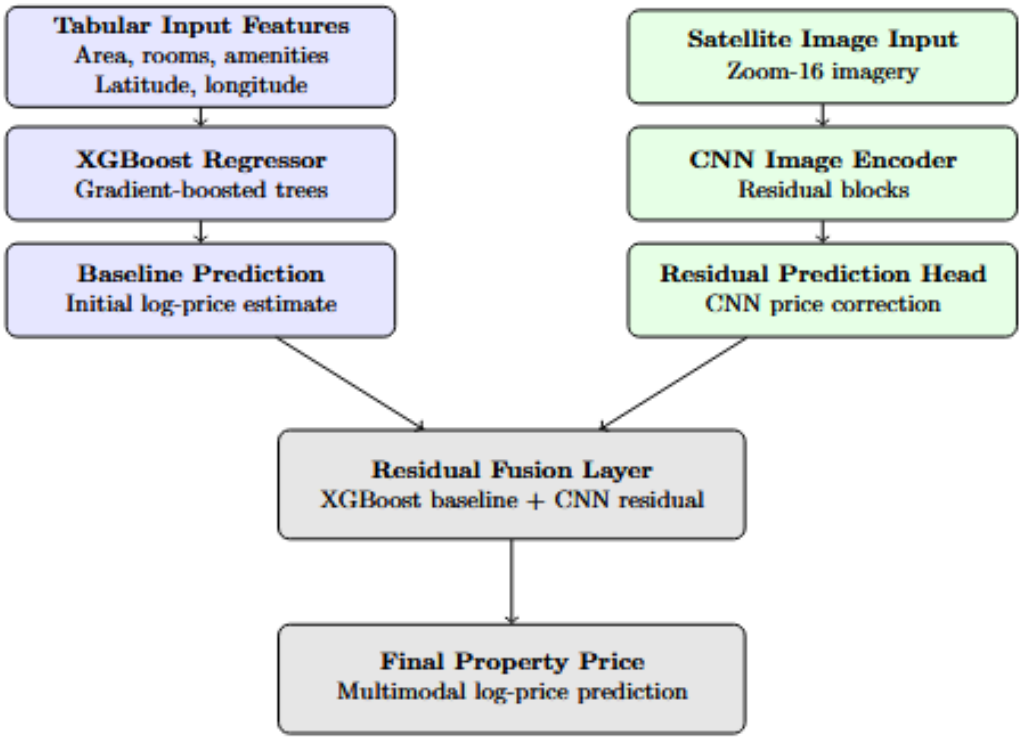

Figure 10: Multimodal residual fusion architecture for property price prediction

In practice, the image branch can process satellite imagery at multiple zoom levels to capture both fine-grained local details and broader neighborhood context.

5 Results and Performance Comparison

In this section, we evaluate and compare the performance of different modeling approaches on the validation set. The goal of this comparison is to assess the impact of incorporating satellite imagery alongside tabular features and to understand how different fusion strategies affect prediction accuracy.

Model	RMSE	R^2
Tabular Data Only (XGBoost)	0.275	0.72
Naive Multimodal Fusion	Higher	Lower
Residual Fusion (Proposed)	**0.257**	**0.76**
Weighted Residual Fusion	0.27	0.74

NOTE: The reported RMSE and R^2 values are evaluated on the validation set rather than the training data. This ensures that the results reflect how well the model performs on unseen data instead of how well it fits the training samples. Consequently, the performance values are more moderate and realistic, avoiding overly optimistic results that can occur when reporting training accuracy.

This table summarizes the performance of different modeling approaches on the validation set. The tabular-only XGBoost model provides a strong baseline using structured features alone.

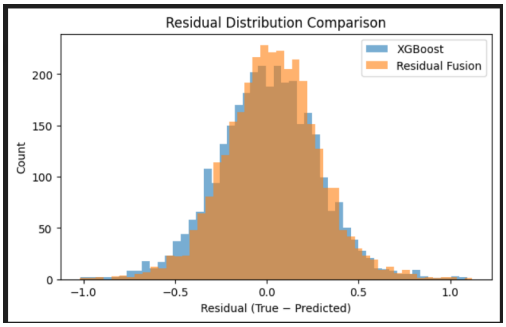

Figure 11: Comparison of residual distributions for tabular-only and residual fusion models

Figure 11 compares the residual distributions of the tabular-only XGBoost model and the residual fusion model. The residual fusion approach shows a tighter concentration of errors around zero, indicating reduced variance and fewer extreme prediction errors. This distributional improvement complements the gains observed in RMSE and R^2.

Overall, the results confirm that satellite imagery contains valuable contextual information for property valuation. However, this information is most effectively utilized when incorporated through a residual learning framework rather than naive feature fusion.

6 Discussion

The experimental results show that satellite imagery can improve property price prediction when visual information is integrated in a structured manner. Tabular features such as property size, quality, and location provide a strong baseline and explain a large portion of price variation.

Naive multimodal fusion fails to improve performance and can degrade accuracy, indicating that unconstrained visual features introduce noise. In contrast, the residual fusion approach enables the image-based model to focus on correcting tabular prediction errors, leading to improved stability and performance.

Grad-CAM analysis reveals that the residual CNN primarily attends to neighborhood-level structures rather than individual buildings, suggesting that satellite imagery contributes complementary contextual information. While the approach has limitations related to image resolution and missing socioeconomic factors, the results demonstrate that residual learning provides an effective and interpretable framework for integrating visual context into property valuation models.