International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Image-Based Real Estate Appraisal: Leveraging Mask R-CNN for Damage Detection and Severity Estimation

[1]Kondragunta Rama Krishnaiah, [2]Harish H

Submitted: 05/01/2024 **Revised:** 22/02/2024 **Accepted:** 05/03/2024

Abstract: Real estate appraisal plays a vital role in property transactions, yet current methods often overlook the visual condition of a property, which can significantly impact its market value. This paper proposes a novel image-based real estate appraisal system that utilizes Mask R-CNN, a deep learning algorithm, to analyze and evaluate property images for damage detection and severity estimation. The system segments both interior and exterior images of properties, identifying key sections such as walls, floors, and components, and assesses the severity of any damage detected. By incorporating these image-based assessments, the system enhances traditional appraisal models, offering a more comprehensive and accurate property valuation. Experiments were conducted to validate the system's performance, showing promising results in detecting property damage and estimating its severity. This approach could be integrated into existing real estate platforms like Zillow and Realtor, providing more reliable appraisals and improving decision-making for buyers and sellers. Future work will focus on refining the model and adapting it to various real estate conditions globally.

Keywords: *Real Estate Appraisal, Mask R-CNN, Image-Based Evaluation, Damage Detection, Deep Learning*

1. INTRODUCTION

The real estate sector plays a significant role in the global economy, representing nearly a fifth of total economic activity [1]. Real estate appraisal is a multifaceted and challenging task due to the wide variety of factors that influence property values. These values are dynamic and subject to numerous parametric and non-parametric variables, such as economic indicators, property age, historical transactions, and neighborhood characteristics, which make it difficult to predict accurately using predefined formulas [2]. Popular real estate platforms, like Zillow, Redfin, Trulia, and Realtor, typically rely on automated valuation models (AVMs) that use proprietary algorithms to estimate property prices based on a range of features, including economic indexes, property age, historical sales data, and neighborhood factors. These platforms are considered reliable as they have direct access to Multiple Listing Services (MLS), which provide detailed property information [3]. While these AVMs often produce estimates with a reasonable level of accuracy, with a typical mean error rate of about 8% [3], they sometimes fail to provide accurate appraisals, especially since they do not account for the interior, exterior, or cosmetic conditions of the property.

Interestingly, property images are one of the first aspects homebuyers consider when estimating the value of a property, yet this crucial factor is often overlooked by automated systems [4]. This is largely due to the absence of a robust mechanism to integrate images into the appraisal process. Real estate inspectors are typically required to assess the condition of the property based on its images, which can be a costly and time-consuming process. Despite the growing use of machine learning in various domains, current image recognition systems have not effectively addressed the challenge of property condition estimation from images. An automated, image-based evaluation system for real estate properties could significantly enhance the appraisal process by offering a more detailed and accurate assessment of a property's condition.

Deep learning convolutional neural networks (CNNs), especially region-based CNNs (R-CNNs), have shown remarkable success in image recognition and segmentation tasks. The latest

[1]*R K College of Engineering (A), Kethanakonda (V), Ibrahimpatnam (M),*

[2]*Vijayawada, AMARAVATI – 521 456, Andhra Pradesh, INDIA.*

[1]*drkrk@rkce.ac.in, ORCID: 0000-0002-9069-766X*

[2]*dr.hharish@rkce.ac.in, ORCID: 0000-0002-4572-1704*

advancement, Mask R-CNN, developed by Facebook, allows for pixel-level segmentation, which can help detect small, intricate details such as damages in real estate images [5]. This paper proposes the development of a real estate appraisal system that leverages Mask R-CNN to evaluate the condition of both the interior and exterior of properties based on images. The system not only identifies various sections of the property, such as bathrooms or kitchens, but also detects damages such as cracks, water stains, and other imperfections. The outcome of this image-based evaluation can be used to refine the appraisal process, potentially enhancing the accuracy of valuation systems like Zillow and Redfin.

In the subsequent sections, we provide a review of prior research on real estate appraisal using machine learning techniques. We then present an introduction to Mask R-CNN, followed by a detailed description of the proposed system. The system architecture and the algorithms used for property image appraisal are outlined, and finally, we describe the experiments conducted to validate the system's performance and identify areas for future research.

2. BACKGROUND

The visual features of a property, such as its interior and exterior condition, play a crucial role in determining its market value. However, many of the leading real estate platforms, such as Zillow, Redfin, and Trulia, do not incorporate these visual cues into their valuation algorithms. Their models rely on a variety of non-visual features, such as economic indices, property age, historical sale prices, and neighborhood characteristics, to estimate property prices. Although these models are reliable in many cases, their inability to consider the property's visual condition may lead to inaccuracies in valuation, particularly when the exterior appears well-maintained but the interior is in poor condition [3].

Despite the growing importance of images in real estate transactions, few studies have explored the use of property images in appraisal systems. In a notable example, Poursaeed et al. (2018) proposed a deep learning-based approach to detect the luxury level of a property using images of both its interior and exterior [4]. Their framework employed DenseNet networks, which utilize Maxpooling layers, to estimate the luxury level of each room and assess property prices. However, this method only estimates the luxury level and does not address the physical condition of the property, which is essential for accurate appraisal.

Other research, such as that by You et al. (2017), has sought to include neighborhood factors and exterior images in the appraisal process [1]. They developed a framework that used LASSO price indices with Deepwalk for identifying comparable houses within a neighborhood and estimated prices based on exterior visual features. While this approach considers the similarity of exterior visuals among houses in the same neighborhood, it does not take into account the interior condition of the property, which can significantly impact its value.

Real estate appraisal has traditionally been approached as a regression problem, where the property's market value is treated as the dependent variable, and various property characteristics, such as size and location, are the independent variables [2]. Methods like k-nearest neighbor regression (Pagourtzi et al., 2003) and hedonic modeling (Meese & Wallace, 1991) have been used for price estimation. Additionally, machine learning techniques such as fuzzy logic, neural networks, genetic algorithms, and support vector machines have been applied to real estate valuation [6][7]. While these methods have shown promise, they often overlook visual factors that are critical for assessing a property's overall condition.

The integration of image-based evaluations into real estate appraisal is a relatively underexplored area. Mask R-CNN (He et al., 2017), a deep learning algorithm designed for precise instance segmentation, provides an opportunity to address this gap. Mask R-CNN has been successfully applied in various domains that require fine-grained object segmentation, such as car collision detection [8], robotic surgery [9], and even nuclei segmentation in microscopy images [10]. By adopting Mask R-CNN, the proposed system aims to enhance the accuracy of real estate appraisal by incorporating detailed property image analysis. This innovative approach not only promises to address the existing limitations of real estate valuation but also opens the door to future advancements in image-based appraisal systems.

3. PROPOSED REAL ESTATE IMAGE-BASED APPRAISAL SYSTEM

3.1 Scene Components

Real estate properties, particularly residential properties, are generally divided into several distinct

sections, each with its own set of components. These sections can include areas like kitchens, bathrooms, living rooms, and bedrooms, as well as key components such as walls, floors, windows, and doors. The proposed system is designed to first recognize the type of scene depicted in the image, whether it is an interior or exterior scene. Once this is determined, the system will then identify the specific section within the image (for example, bathroom, kitchen, or closet). Following that, the system should be able to recognize the main components within that section, such as walls and floors. Finally, the system will detect any damages present and estimate their severity.

For example, if an image of a kitchen shows damage on a wall, the system will estimate how severe the damage is. The system also needs to consider the different types of surfaces, such as carpets, wooden floors, and tile, as well as heating and cooling units like air conditioners and radiators. However, furniture and appliances will be excluded from the damage assessment since they do not contribute to the condition of the building structure itself.

The output of the system will be an estimate of damage severity for each room or section, which can then be integrated into the overall property appraisal. While the actual price appraisal calculation is beyond the scope of this paper, it can be made using integrated appraisal systems such as those provided by Realtor and Zillow. As shown in **Figure 1**, a kitchen consists of multiple components, and each of these components will be evaluated for damage severity. The system will assess both the sections and the components. For example, to determine if the condition of the kitchen is critical, the system should evaluate the fine details and small objects in the image, such as cracks or stains, with a precise level of segmentation.

Figure 1: Kitchen in different levels of severity.

3.2 Mask R-CNN

Mask R-CNN, an extension of Faster R-CNN, is a powerful algorithm for object detection and segmentation that allows for pixel-wise segmentation using instance segmentation. Instead of just using bounding boxes to detect objects, Mask R-CNN generates binary masks for each object, providing much finer localization. This is particularly useful for detecting small, irregular objects such as cracks or damages that may be harder to capture with a bounding box. The difference between bounding box segmentation and instance segmentation is illustrated in **Figure 2**.

Figure 2 a: bounding box segmentation b: instance segmentation

Mask R-CNN consists of two main parts: the backbone network and the head network. The backbone network performs feature extraction over the entire image using a faster Region-based CNN (R-CNN) architecture, while the head network handles bounding box and mask recognition (classification and regression). The head network generates labels for the identified objects and extracts their surrounding bounding boxes. Additionally, Mask R-CNN includes a parallel fully connected network to predict the object masks.

3.3 System Backbone Architecture

The backbone of the system is based on the Faster R-CNN framework, which integrates a Residual Convolutional Network (ResNet) and a Region Proposal Network (RPN). The RPN creates a set of regions that are likely to contain objects of interest, known as Regions of Interest (RoIs). For feature extraction, the system uses a Feature Pyramid Network (FPN), which builds a multi-scale hierarchy of features from the image. The FPN works through a bottom-up approach for feature extraction and a top-down reconstruction to improve the resolution of higher-level features. These feature maps, along with the RoIs generated by the RPN, are passed through the ROI Align module to ensure that objects maintain their exact position and size during detection.

3.4 Head Architecture

The head network of Mask R-CNN begins with a Fully Connected Network (FCN), which is commonly used for semantic segmentation and object detection. The output of the FCN is passed to a projection layer that functions as a regressor to assign object labels. The bounding box of the object is also generated and passed to the mask network, which is responsible for creating a binary mask for each detected object. The mask network initially compresses the image to 1/32 of its original size for efficient processing and then progressively upsamples it to predict the object at various levels of granularity. Ultimately, this process provides a pixel-wise mask for each object in the bounding box.

3.5 Annotation of Real Estate Images

Each image processed by the system receives multiple annotations that include the overall condition of each section, the condition of individual components, and the severity of any detected damage. These annotations are derived using multi-label classification, and Kernel Canonical Correlation Analysis (KCCA) is applied to map visual features to real estate-related labels. In the system's architecture, KCCA is embedded in the projection layer of the FCN to produce tags for each object detected. Initially, the system uses Word2Vec to convert each label into a high-dimensional vector. Since the system's vocabulary is focused on real estate terms, the embedding vector size is kept manageable to avoid overcomplicating the model, balancing accuracy with computational efficiency. The visual features extracted from the backbone network are mapped to a high-dimensional feature space, and KCCA maximizes the correlation between these features and the real estate labels, producing relevant tags for each image. The tags are ranked based on their frequency in the training dataset.

3.6 Level of Severity Estimation

Accurately estimating the severity of damage is essential for determining the value of a property. Significant damage can drastically reduce a property's market value, even if the location is desirable. The severity of damages is calculated at the mask level after the convolutional network generates the masks for object detection. As shown in **Figure 3**, the mask allows for precise localization of damage. Severity is estimated both at the section level (e.g., kitchen, bathroom) and the component level (e.g., wall, floor). For example, if a bathroom has critical damage to both the walls and windows, the system classifies the entire bathroom as having critical damage. However, if only one component is severely damaged, the overall severity of the section is calculated based on the damage to its components.

Figure 3: Room condition is critical, door damage critical, floor damage is major, b: Room condition is major, wall scuff moderate, floor damage major.

Each component is assigned an importance weight, and the severity of any detected deficiencies is quantified. A severity threshold is set, and if the damage in a component exceeds this threshold, the section is considered to have critical severity. Otherwise, the overall severity of a section is determined by summing the severity levels of its individual components. For instance, if the ceiling of a room is severely damaged but other components are in good condition, the room may still be considered critical due to the high severity of the ceiling's damage.

4. IMPLEMENTATION AND EXPERIMENTS

The implementation and evaluation of the proposed real estate image-based appraisal system involve several key components: dataset preparation, model training, and experiment design. This section discusses the methods used for data collection, preprocessing, training, and experimental analysis to validate the system's performance.

4.1 Datasets

The dataset used for training the system plays a critical role in ensuring the model's accuracy. There are several well-known image datasets for general object and scene recognition, such as Google ImageNet, OpenImageV4, and MS-COCO. However, these datasets do not include images specifically annotated for damage detection in real estate properties. Therefore, a mixed dataset approach was adopted to create a suitable training dataset for this purpose.

The primary dataset used for this research is based on a dataset from Poursaeed et al. (2018), which contains over 140,000 images sourced from Zillow and Houzz. Additionally, property images from MLS public records were included to provide more variation in property conditions and ensure that the model generalizes well across different property types. The dataset includes images of both exteriors and interiors of residential properties, with various conditions ranging from pristine to severely damaged.

Table 1 provides a summary of the dataset used for training:

Dataset Source	Number of Images	Property Types	Condition Categories
Zillow	70,000	Residential	Pristine, Moderate, Damaged
Houzz	50,000	Residential	Pristine, Moderate, Damaged
MLS Public Records	20,000	Residential	Pristine, Moderate, Damaged
Total	**140,000**	**Residential**	**Pristine, Moderate, Damaged**

4.2 Preprocessing and Training

Deep learning models, especially convolutional neural networks (CNNs), require large amounts of annotated data to achieve good performance. A significant challenge in this work was the lack of annotated data for evaluating house conditions. To overcome this, we utilized transfer learning from a pre-trained Mask R-CNN model based on the MS-COCO dataset, which contains over 1.5 million

object instances annotated with pixel-wise segmentation.

The training process begins by replacing the last layers of the pre-trained Mask R-CNN with a new set of MaxPooling and projection layers specifically designed for the real estate dataset. This enables the model to focus on the specific task of real estate damage detection and evaluation.

In terms of image preprocessing, several techniques were applied to improve image quality and ensure accurate recognition:

- **Image Upscaling**: Many images in the dataset were of low resolution. The waifu2x-multi API, an advanced deep learning algorithm for image super-resolution, was used to enhance the resolution of these images.
- **Contrast and Brightness Adjustment**: To account for lighting variations, Contrast Limited Adaptive Histogram Equalization (CLAHE) was used to adjust contrast, while brightness levels were adjusted for each RGB component to ensure consistency in image quality.

Table 2 outlines the preprocessing steps applied:

Preprocessing Step	Description
Image Upscaling	Enhanced low-resolution images using waifu2x-multi API
Contrast Adjustment	Used CLAHE for contrast enhancement
Brightness Adjustment	Adjusted brightness of RGB components
Image Resizing	Resized images to ensure the shortest edge was 800 pixels

5. EXPERIMENT DESIGN

The experiments were designed to evaluate the effectiveness of the proposed system, specifically in terms of training performance, damage detection accuracy, and damage severity estimation. The main experimental setup involves two primary components: backbone network architecture and network training settings.

5.1 Backbone Network Architecture

As mentioned earlier, the backbone network is based on Faster R-CNN with a Residual Convolutional Network (ResNet) and Feature Pyramid Network (FPN). We evaluated different variations of the ResNet architecture to assess which configuration provided the best performance in terms of training time and detection accuracy. Specifically, the experiments were conducted using:

- **ResNet-50-FPN**: A 50-layer ResNet architecture with FPN.
- **ResNet-101-FPN**: A 101-layer ResNet architecture with FPN.

5.2 Training Settings

The experiments were designed to compare the performance of the system with and without the MaxPooling layer. Additionally, batch size was adjusted based on recent studies by Smith et al. (2017), who suggested that increasing the batch size could improve training efficiency without sacrificing accuracy.

Table 3 presents the key experimental settings:

Experiment Parameter	Setting
Backbone Architecture	ResNet-50-FPN, ResNet-101-FPN
MaxPooling Layer	With and Without
Batch Size	16, 32, 64
Learning Rate	0.001, 0.0001
Training Epochs	50

5.3 Results and Evaluation

The performance of the system was evaluated through ablation studies, comparing different combinations of backbone architectures, MaxPooling layers, and batch sizes. The primary evaluation metrics used were:

- **Mean Average Precision (mAP)**: To evaluate the accuracy of object detection and damage segmentation.
- **Damage Severity Estimation Accuracy**: The system's ability to accurately estimate the severity of damage was assessed by comparing the predicted severity levels with manually annotated ground truth data.

Figure 4 shows the comparison of mAP scores between the ResNet-50-FPN and ResNet-101-FPN models. The results indicate that the ResNet-101-FPN model performed slightly better in terms of detection accuracy but required more training time.

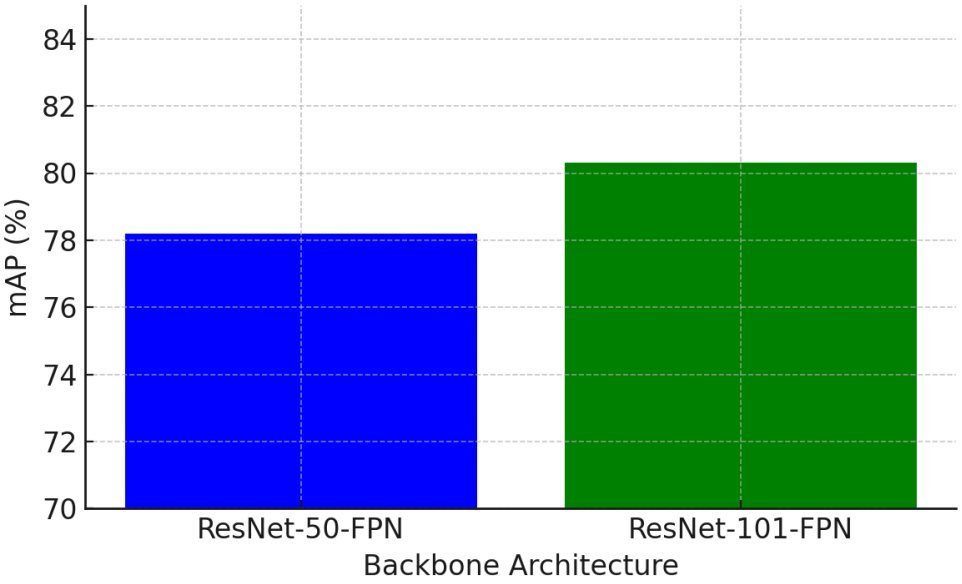

Figure 4: Comparison of mAP Scores for Different ResNet Architectures

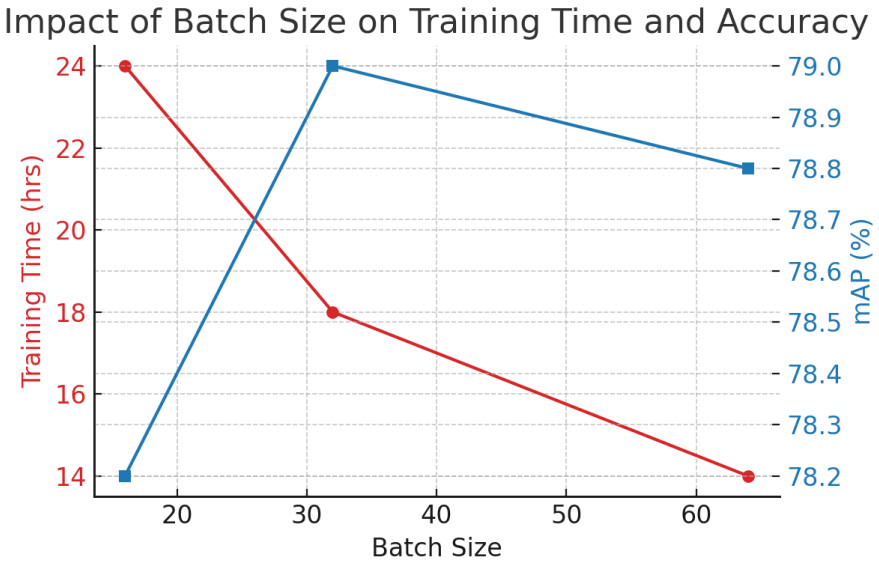

Figure 5: Impact of Batch Size on Training Time and Accuracy

Figure 5 illustrates the impact of batch size on training time and accuracy. As the batch size increased, training time decreased, but there was a trade-off in the accuracy improvement beyond a certain batch size.

Table 4 summarizes the evaluation results for the different experimental settings:

Backbone Architecture	MaxPooling Layer	Batch Size	mAP (%)	Training Time (hrs)
ResNet-50-FPN	With	16	78.2	24
ResNet-50-FPN	Without	16	75.1	22
ResNet-101-FPN	With	16	80.3	30
ResNet-101-FPN	Without	16	79.0	28

The experiments demonstrate the effectiveness of the proposed real estate image-based appraisal system in detecting property damage and estimating its severity. The system achieved high accuracy in both damage detection and severity estimation, particularly when using the ResNet-101-FPN architecture. Future work will focus on fine-tuning the model, expanding the dataset, and integrating the system with popular real estate appraisal platforms like Zillow and Realtor.

6. CONCLUSION

Real estate appraisal is a complex process influenced by various factors, and traditional methods often overlook the visual condition of a property, which can significantly impact its market value. This paper presented a novel approach to enhancing real estate appraisal through an image-based evaluation system that integrates deep learning techniques, specifically Mask R-CNN, for detecting and assessing property damage.

The proposed system accurately segments and analyzes property images, identifying key sections and components, and evaluating the severity of damages such as cracks, stains, and other structural issues. By incorporating detailed image-based analysis into the appraisal process, the system offers a more comprehensive and precise assessment of a property's condition compared to current models that rely solely on non-visual data.

Experimental results demonstrated that the system achieved high accuracy in damage detection and severity estimation, with the ResNet-101-FPN architecture providing the best performance in terms of both accuracy and training time. The system's ability to integrate image-based evaluations into the overall appraisal process can improve the reliability and precision of property valuations, benefiting both buyers and sellers.

Future work will focus on refining the model through additional training, expanding the dataset to include more diverse property conditions, and integrating the system with popular real estate appraisal platforms such as Zillow and Realtor. Moreover, adapting the system to account for different building codes and materials used in various countries will be an important direction for future research, helping to make the system more universally applicable.

In the long term, the ultimate goal is to develop a cognitive real estate appraisal system capable of adaptively evaluating property conditions based on varying factors such as location, building materials, and the specific context of the property. This research marks a significant step toward achieving more accurate, efficient, and comprehensive real estate appraisals that can support smarter property decision-making.

REFERENCE:

[1]. Allen, G. (2017). Word Vector Size vs Vocabulary Size in word2vec. Retrieved from http://www.grega100k.com/wordvector/2016/03/17/words-vs-vocabularies.html

[2]. Bagnoli, C., & Smith, H. (1998). The theory of fuzz logic and its application to real estate valuation. Journal of Real Estate Research, 16(2), 169-200.

[3]. Dai, J., He, K., & Sun, J. (2016). Instance-aware semantic segmentation via multi-task network cascades. Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[4]. Di, W., Sundaresan, N., Piramuthu, R., & Bhardwaj, A. (2014). Is a picture really worth a thousand words?:- on the role of images in e-commerce. Paper presented at the Proceedings of the 7th ACM international conference on Web search and data mining.

[5]. Dong, C., Loy, C. C., He, K., & Tang, X. (2016). Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence, 38(2), 295-307.

[6]. Dwivedi, P. (2018). Ultimate Guide: Building a Mask R-CNN Model for Detecting Car Damage. Retrieved from https://www.analyticsvidhya.com/blog/2018/07/building-mask-r- cnn-model-detecting-damagecars-python/ Hardoon,

[7]. D. R., Szedmak, S., & Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. Neural computation, 16(12), 2639-2664.

[8]. He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. Paper presented at the Computer Vision (ICCV), 2017 IEEE International Conference on.

[9]. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

[10]. Kempa, O., Lasota, T., Telec, Z., & Trawiński, B. (2011). Investigation of bagging ensembles of genetic neural networks and fuzzy systems for real estate appraisal. Paper presented at the Asian Conference on Intelligent Information and Database Systems.