Multi-modal Apartment Appraisal: Exploration of Deeper Architectures and Geo-temporal Normalisation Scheme

Beāte Desmitniece, Patryk Kuchta, Aleksander Kirsten

Abstract

Earlier efforts to automatically perform property appraisal by using multi-modal input features have relied on relatively shallow neural network architectures. This study proposes a new architecture combining a feed-forward and DenseNet convolutional neural networks, whilst applying transfer learning and advanced regularisation techniques to tackle the task more effectively. Additionally, to account for the property market diversity across various locations and sales periods, we propose a geo-temporally normalised objective (loss) function. Lastly, our research is the first to experiment with the incorporation of point-of-interest and transport maps in the input feature set. All experiments were conducted on a partially self-collected Latvian apartment sales dataset. On the task of apartment value prediction, the introduced architecture trained using the proposed geo-temporal normalisation loss functions achieves a 0.7287 R^2 score, outperforming previously employed methods from the real estate literature.

1. Introduction

Property appraisal is the process of assessing the current market value of a given real estate. Unfortunately, the appraisal procedure is expensive and time-consuming as it is typically done by a manual site inspection. This challenge could be effectively addressed by employing automatic real estate valuation techniques, ensuring rapid and inexpensive property valuation. Moreover, the use of automatic property appraisal tools would ensure consistency and reduce variation, as the decision of human valuers can be influenced by the knowledge of previous value estimates or by the presence of the client (Baum et al., 2021). Additionally, an automatic appraisal can be used by governmental institutions to detect tax fraud by comparing the model-predicted price with the listed property sale price.

This research focuses on developing a deep-learning property appraisal model that predicts the price based on the property's quantifiable attributes and the spatial visual information of the surrounding area presented through various maps. More specifically, we propose a novel multi-modal architecture, combining DenseNet (Huang et al., 2017) and feed-forward architectures that accommodate quantitative tabular and image data input types with improved regularisation techniques.

Additionally, based on the findings of Yang et al. (2020) and Vivian W. Y. Tam & Ma (2022), which demonstrate a correlation between property value and the proximity of transportation and neighbourhood amenities, we utilise transportation map and point-of-interest (POI) map as input values of the price prediction model.

Lastly, to account for the high variance in property prices in various locations and sale times, we propose a novel objective loss function that abstracts away from the market dynamics, allowing for accurate price prediction.

This study aims to:

1. Develop a new geo-temporally normalised loss function for property valuation and compare its performance to conventional MSE objective function.
2. Investigate the effect on model performance when point-of-interest maps and transportation maps are added to the input feature set alongside satellite imagery.
3. Propose and evaluate a deep multi-modal network for property price prediction, which combines DenseNet and feed-forward architectures with improved regularisation techniques. Compare its performance against baselines from real estate valuation literature.

All experiments are conducted using open-source Latvian government data of apartment sales over the past three decades, along with additional self-collected map images.

2. Related Work

Various research studies have been conducted in the domain of housing value estimation with multi-modal deep neural architectures utilising spatial visual images of the surrounding area, each modelling the complex dynamics of real estate markets.

Bin et al. (2019) propose a network architecture inspired by attention mechanisms for predicting house prices in Los Angeles, USA. They utilise a three-layer convolutional neural network (CNN) to extract features from street map images. These features, along with house attributes, are then processed through attention blocks and a feed-forward network. The resulting encodings serve as features to train a boosted regression tree, which is used for estimating house values. Their research supports the claim that the incorporation of spatial information improves the performance of the real estate valuation model.

Alternatively, Bency et al. (2017) collect satellite images at various scales and apply fine-tuned Inception v3 (Szegedy et al., 2016) CNNs for feature extraction. The image features are combined with house attributes and nearby point-of-interest locations and parsed through a feed-forward network to predict the house price. The researchers experimented with data from London, Birmingham, and Liverpool and concluded that satellite image features and POI attributes contribute to a significant improvement in model performance. We take inspiration from the proposed method and utilise point-of-interest data for property price prediction. However, we incorporate POI data into the property price prediction model as image data. Additionally, our study aims to develop a general model applicable to the entire housing market of a country, rather than having separate models for each city tested, as done by Bency et al. (2017).

The research of Azizi & Rudnytskyi (2022) proposes a multi-modal architecture where property attributes are parsed through a feed-forward network and fed into another feed-forward network along with three-layer CNN features extracted from a satellite image. Drawing inspiration from their method, we aim to create a single model applicable to the entire housing market of a country, similar to the study's approach, which developed a housing price estimation model for the whole of Switzerland. We use their proposed method as a baseline and develop our model by modifying their network architecture. Specifically, we utilise a deeper CNN and replace late fusion with intermediate fusion, arguing that with such modification, the model might learn more complex associations.

It must be highlighted that the aforementioned studies use the property's coordinates as one of their inputs, whereas we intend to make our model location agnostic, preventing the model from making associations with expensive or moderately priced areas.

In their work Law et al. (2019) utilise VGG CNNs (Simonyan & Zisserman, 2015) to extract features from street view and satellite images and parse them through a feed-forward neural network, alongside house attributes encoded using a feed-forward network. The researchers test their proposed method on properties in London. Inspired by the proposed architecture, our study employs intermediate fusion and further explores the impact of model depth on the correctness of property price predictions. Hence, Law et al. (2019)'s work is selected as our second baseline.

Our research highlights and addresses the research gap by developing a country-level apartment prediction model independent of the property location and time of sale by proposing a deep multi-modal architecture with a greater level of regularisation.

3. Dataset and Task

3.1. Quantitative Data

The experiments are conducted using a publicly available real estate market database created and maintained by the State Land Service of Latvia[1]. The data consists of all real estate transactions registered in Latvia's digital land register from 1998 until 2024. We extract a subset of the original dataset, comprising only apartment sales. For every apartment, the extracted feature set includes the transaction and property identifier, address, closest city or town (referred to as "town" in this paper), date of transaction, property price, building material, number of floors in the apartment building, years when the building received a certificate of occupancy[2], floor area of the apartment, floors on which the apartment is located, and indicators of whether the apartment building has any public[3] or commercial space. We clean the data by removing duplicate entries, any entries with missing values, and properties where the room count is less than one.

The building material is encoded with 1-hot encoding in four types: brick, wood, concrete, and others. Based on information on the years the building obtained its certificate of occupancy, two features are included for each property, illustrating the year the building was built and the year it had its most recent refurbishment. Furthermore, we calculate the number of floors an apartment has. Additional features for each property are introduced to account for discrepancies in property values across different years and various towns of Latvia due to historical price trends and differences between local property markets. We calculate the property price per m^2, the average price per year per town per m^2, the normalised property price per m^2 (see Section 4.3) and the floor count within the flat. If the property price per m^2 is less than 10 euros, the transaction is discarded.

We utilise Bing Locations API[4] to obtain the longitude and latitude of the property based on the apartment address. Additionally, we acquire the geographical centre coordinates of all cities in Latvia from the Latvian Wikipedia[5]. Using the collected locations, we calculate each property's distance to its respective town centre and add it to the feature set. Assuming every site in Latvia is within a 40 km distance from a town, properties with a distance measure greater than 40 km are discarded, indicating an incorrect location response from the API. Based on the analysis of API responses, we approximate that above 20% of the property coordinates were faulty.

[1] data.gov.lv/dati/lv/dataset/nekustama-ipasuma-tirgus-datu-bazes-atvertie-dati

[2] Certificate of Occupancy is a document certifying that a building complies with build standards and is suitable for occupancy.

[3] Community facilities, such as educational, religious, health and social institutions.

[4] learn.microsoft.com/en-us/bingmaps/rest-services/locations/find-a-location-by-address

[5] lv.wikipedia.org

		Abbr.
Quantitative Features	Date of sale	T
	Price, EUR	
	Price per m^2, EUR	
	Normalised price per m^2, EUR	
	Average town price per m^2, EUR	
	Has commercial space?	
	Has public space?	
	Floors in building	
	Building built year	
	Building last renovation year	
	Material of building (brick, wood, concrete, other)	
	Floor number	
	Number of floors in apartment	
	Number of rooms	
	Floor area, m^2	
	Distance to centre	
Images	80 m radius satellite image (250x250)	S_{80}
	300 m radius satellite image (250x250)	S_{300}
	600 m radius transport map image (250x250)	TM
	600 m radius Tracestrack Topo map image (250x250)	POI

Table 1. The set of all available features in our dataset.

To account for the various market dynamics, the properties can be grouped into categories such as luxury housing or rural housing. It is appropriate to cluster the properties with similar characteristics together before performing automated outlier detection (Ozer & Okan Sakar, 2022), regardless of their price. Principal Component Analysis (PCA) is performed on all features, excluding the price-related descriptors. The first five principal components are retained, and K-means clustering with 10 clusters is performed. In every cluster, all data points outside the $1.5 * IQR$ (interquartile range) of the property price per m^2 are considered outliers and are discarded.

3.2. Image Data

Furthermore, satellites.pro and openstreetmap.org map services are utilised to obtain four 250x250 coloured images for every property. These images encompass two satellite views, with one covering an 80 m radius surrounding the property and the other extending to a 300 m radius. Additionally, the image set includes a 600 m radius transport map illustrating public transport routes and a Tracestrack Topo map displaying tags of nearby POI. To optimize the fetching process and the size of the dataset, properties within 10-meter proximity have been merged, and they share the satellite and OSM images.

The cleaned dataset consists of 149,997 property transactions with the respective features and images. The dataset is split into training (85%), validation (5%) and two testing sets (each 5%). The first test set consists of randomly sampled properties, whereas the other test set consists of all properties located in the towns of Rēzekne, Ogre, and Valmiera. It has been ensured that no geographically merged location images are 'shared' between sets. We intend to use the unseen town set to evaluate the model's ability to forecast the property price in a previously unseen environment. The full feature set is illustrated in Table 1.

4. Methodology

4.1. Baseline models

To assess the viability of our proposed model, we assess its performance against network architectures from multimodal property valuation literature. In this subsection, we introduce the architectures of the selected baselines.

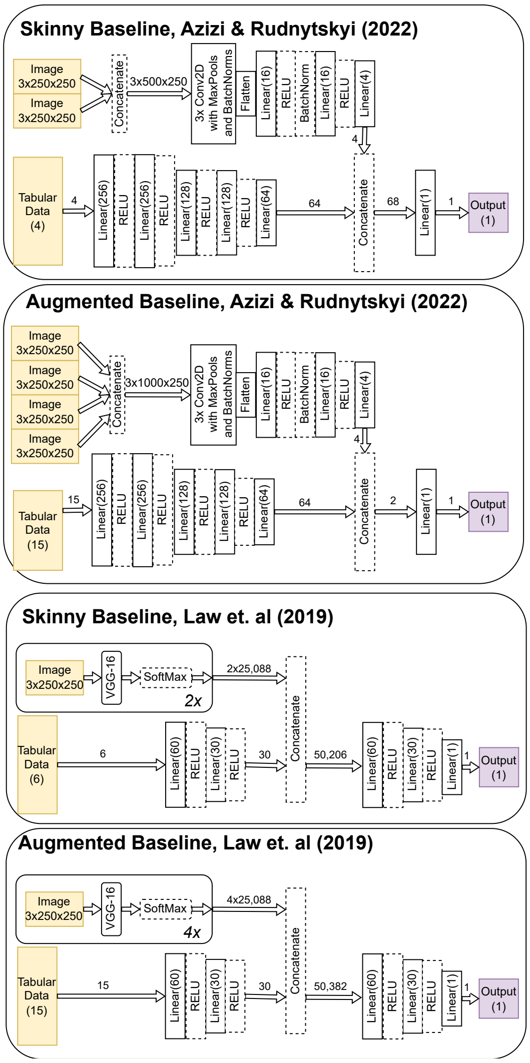

Figure 1. Diagram showing our adaptations of the architectures of Azizi & Rudnytskyi (2022) and Law et al. (2019). Those will serve as our baselines. Details about the architecture of VGG-16 are covered by Simonyan & Zisserman (2015).

The first baseline is a CNN architecture composed of three convolutional layers that take as input concatenated images and produce an output which gets flattened and is processed by 2 linear transformations (Azizi & Rudnytskyi, 2022). The tabular input of the network is processed by 5 linear transformations. The results of each of the parts of the model are then concatenated together and passed through a final linear layer, highlighting the use of late fusion. The activation function after each transformation is the ReLU function. Additionally for regularization dropout (Srivas-

Baseline	Original features	Proxy features	Abbr.
1.	Living space, m^2	Floor Area, m^2	T_S
	Number of rooms	Number of rooms	
	Longitude	Distance	
	Latitude		
	Satellite image	80 m radius satellite image	S_{80}
2.	Type	*Has public space?*	T_S
		Has commercial space?	
	Year	Date of sale	
	Age	Building built year	
	Size	Floor Area, m^2	
	Beds	Number of rooms	
	Park	*600 m radius Tracestrack Topo map image*	POI
	Shop		
	Gravity		
	Satellite image	300 m radius satellite image	S_{300}
	Streetview	-	-

Table 2. The quantitative features for the *skinny baseline*. The *italics* in the table highlight cases where the proxy relation was particularly weak. 1. Azizi & Rudnytskyi (2022); 2. Law et al. (2019)

tava et al., 2014) and batch normalization (Ioffe & Szegedy, 2015) were used at various points. A more detailed outline of the model is illustrated in Figure 1.

Additionally, as the second baseline, we choose an architecture proposed by Law et al. (2019), who utilised a CNN architecture inspired by VGG-16 network (Simonyan & Zisserman, 2015) to parse images. Contrary to Azizi & Rudnytskyi (2022) they did not concatenate the images, but rather used a separate CNN for each of the images. The outputs of these CNNs were flattened and directly concatenated with the output of tabular transforms. The tabular input of the network was processed by 2 linear transformations. The result after fusing the outputs was then passed through 2 further linear transformations, indicating intermediate fusion. The activation function after each transformation was the ReLU function. A more detailed outline of the model is illustrated in Figure 1.

We intend to train and compare both of these architectures in two different scenarios with varying input features. First, both baselines will be trained using the features, the authors have used in their original research. Further, models will be provided with the entire set of features available in our dataset, and training will be conducted. Unfortunately, our dataset does not have the exact features that the authors of the discussed papers had used in their research. Hence, we will be using the proxies closest to these features that are available in our data set (see Table 2). The models using the full set of features are referred to as *augmented* baselines, while those utilizing the limited set are referred to as *skinny* baselines.

4.2. Proposed Architecture

We propose a multi-modal architecture that combines image features with tabular data using intermediate fusion and employs transfer learning. The network encompasses the DenseNet-121 architecture (Huang et al., 2017) for image processing followed by an affine layer. The outputs of the CNN networks are concatenated with the parsed tabular data and processed through feed-forward blocks to make a price estimation. The multi-modal architecture is illustrated in Figure 2. Initially, we train separate models that utilise only tabular data and image data, after which we transfer the pre-trained weights to the multi-modal architecture. In this section, we explain the rationale and the development process of the proposed network.

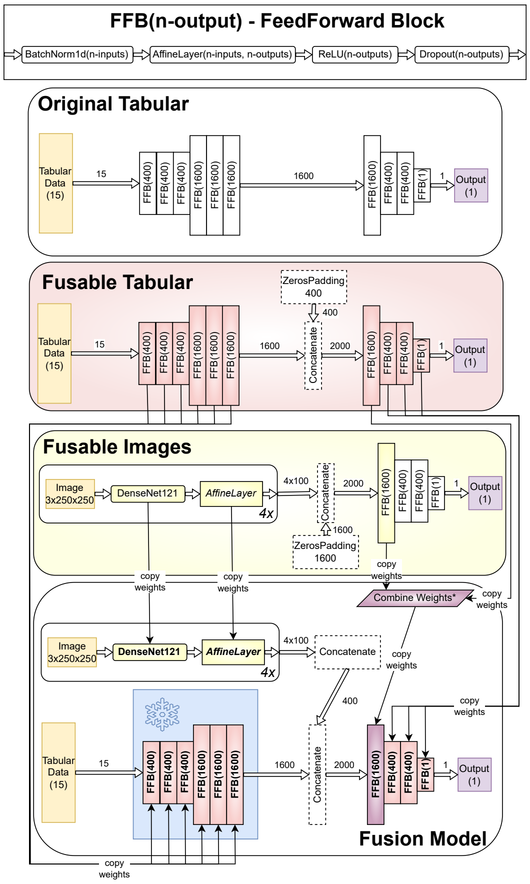

Figure 2. Diagram showing the architecture of the proposed model in various development stages. Layers in **bold** are initialized with pre-trained parameters, and the layers grouped with a snowflake are *frozen* (suppressed weight updates) for the duration of training. Details about the architecture of DenseNet-121 are covered by Huang et al. (2017).

We began the model development by designing an architecture that utilises only tabular data for apartment price prediction. Inspired by the work done by Jiang et al. (2022) we utilised feed-forward blocks which are composed of batch normalization (Ioffe & Szegedy, 2015), affine transformation, rectified linear unit (ReLU) and finally, a dropout layer (Srivastava et al., 2014). The depth of the model was tuned to be 10 layers. This decision was based on experimentation with various depths, and 10 layer architecture was the shallowest architecture that could fit the training data extremely well (*Not Densely Connected** in Table 3). The width of

Experiment		Epoch	Train MSE	Valid R^2
Densely Connected		191	0.1584	**0.5849**
Not Densely Connected*		200	**0.0331**	0.5480
Dropout	0.2	200	0.2236	0.7555
	0.5×10^{-2}	93	0.2000	0.7977
	2.5×10^{-2}	83	0.1895	**0.8182**
	1×10^{-3}	42	**0.1524**	0.7797
L2 Decay	0.1	157	0.1881	0.8221
	0.5×10^{-2}	189	0.1843	**0.8265**
	2.5×10^{-2}	194	**0.1819**	0.8232
LR	1×10^{-4}	93	**0.1810**	0.8063
	1×10^{-6}	189	0.1843	**0.8265**
	1×10^{-8}	200	0.7947	0.1312

Table 3. The best epoch metrics of the most important experiments for tabular architecture tuning. Entries in **bold** are the best in their comparison group. 'Epoch' indicates the number of epochs where these results appeared. Valid - Validation; LR - Learning rate; Not Densely Connected* - No regularization applied

the hidden dimensions was tuned in the same experiment resulting in the following hidden dimensions: [400, 400, 400, 1600, 1600, 1600, 1600, 400, 400] (*Original Tabular* in Figure 2).

Further, we built a *Fusable Tabular* model, which would be trained on tabular data only, but additionally appends padding to the layer where in the future the parameters from the image layers will come in. This peculiar difference between *Original Tabular* and *Fusable Tabular* will allow us to transfer all weights from the *Fusable Tabular* Model into the final *Fusion Model* which will replace the padding with the output of the image layers. The point of fusion was selected to be 4 feed-forward blocks before the output, inspired by similar papers fusing image data at an intermediate point (Law et al., 2019).

Furthermore, to obtain the pre-trained weights of the parameters handling images, we have trained a *Fusable Image* Model consisting of multiple DenseNet-121 networks and an affine layer. The outputs of those networks are concatenated and passed to all the post-fusion feed-forward blocks. The parameters expected from the tabular part of the model are replaced with a padding of zeros.

All the pre-trained parameters related to images, including 400 rows of the first post-fusion linear transform, have been copied into the final *Fusion Model*, and the tabular parameters have been copied from the *Fusable Tabular Model*. The use of transfer learning is inspired by Bency et al. (2017) who have shown model performance improvement for the task of property value estimation with multi-modal attributes.

We have experimented with encompassing dense connections for the feed-forward blocks similar to how it was done in the original paper (Jiang et al., 2022). Unfortunately, those experiments proved unfruitful, as we were unable to find a satisfactory balance between regularization and network capacity, leading to excess underfitting or overfitting in every experiment. Table 3 shows the best densely connected network (*Densely Connected*) aided by a substantial amount of regularization and one without any regularization and no-dense connections (*Not Densely Connected**). Even though DenseNet achieved a better validation R^2 result, it was unsatisfactory given the amount of regularization required to reach this point. On the other hand, dropping the dense connections allowed the model to get very close in terms of validation R^2 whilst not requiring regularization of any kind, pointing to the fact that this model has much more potential.

Learning rate, dropout and L2 regularization hyperparameter tuning for the *Fusion Model* were performed by incrementally adjusting the values based on the previous sweep outcomes to maximise the best validation R^2 metric after 200 epochs, resulting in a choice of 10^{-6}, 5×10^{-2} and 2.5×10^{-2}, respectively. We have included a selection of the results from the hyperparameter tuning in Table 3. The optimizer for this model is ADAM (Kingma & Ba, 2017).

4.3. Objective function

We propose a novel geo-temporally normalised objective function, to compensate for the high price variation in property value estimation. This subsection explains the rationale for its necessity, examines earlier uses of normalisation functions in the literature on property price estimation, and explains the intricacies of the proposed objective function

4.3.1. Rationale for Geo-Temporal Normalisation

The analysis of the Latvian housing market has highlighted an extreme variety of properties in various towns. More specifically, the property price distribution is skewed towards central Latvian towns, where the properties are significantly more expensive than the rest of the country (see Figure 3). Additionally, as the property prices have been collected over a vast time range, they exhibit a high discrepancy where the average property price in the country has risen significantly over the years (see Figure 4). These observations highlight the need for property price adjustments that account for the variance in the location and purchase time when developing a nation-level real estate evaluation model.

4.3.2. Rationale for Price Normalisation

By using typical optimization functions, such as MSE, on high variance target prices, the model will prioritise the modelling of expensive properties over moderately priced ones, as the errors of such targets will have a greater influence on model parameters. This introduces an inherent bias where more expensive property prices will be estimated more accurately than low-cost real estate. Such inaccuracy presents negative implications and the necessity to revert to manual property valuation. However, manual real estate appraisal fees do not directly scale with the value of the property. Hence, the cost of the inspection is more pronounced at the more affordable end of the property market.

One could mitigate this issue by having an automatic property valuation tool capable of accurately assessing the price of moderately and highly-priced properties, with equal importance assigned to both categories.

Such observations imply the need to shift the emphasis away from modelling the high-end real estate market to make it more applicable for properties of the whole price spectrum.

A substantial amount of prior research in property value prediction has identified this issue, and to dampen its effects a logarithm formula has been devised to reduce the prices on the higher-end extremum (Law et al., 2019; Chen et al., 2022; Azizi & Rudnytskyi, 2022; Bin et al., 2019). To normalise a price (p_i) of sale i under this scheme one can use the Equation 1 to derive the normalised price (y_i):

$$y_i = \ln(p_i) \tag{1}$$

The inverse of this function (Equation 2) can be used to convert the output of the model ($\hat{y}_i$) to the value prediction ($\hat{p}_i$).

$$\hat{p}_i = 2^{\hat{y}_i} \tag{2}$$

This adjusted price ($\hat{y}_i$) can then be used in the standard mean squared error loss function (Equation 3).

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{3}$$

4.3.3. Geo-Temporally Normalised Objective Function

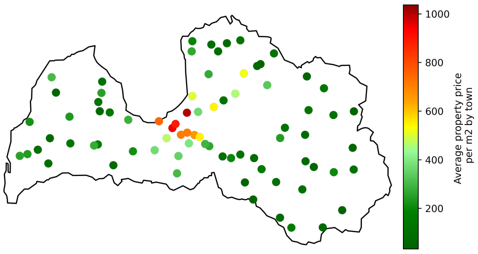

Figure 3. Thematic map of average apartment prices per m^2 in towns of Latvia from 1998 to 2023.

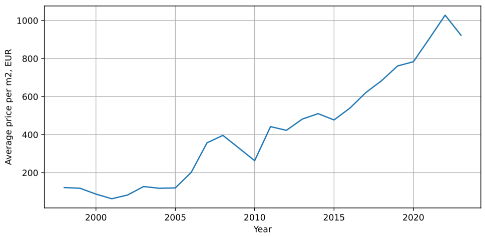

Figure 4. Average apartment price per m^2 in Latvia from 1998 to 2023

We propose an objective function with the intent to remove the inherent bias and provide enhanced predictions for all market tiers, including those categorized as budget-friendly, moderate, and high-end. Additionally, the proposed objective function abstracts away from the location of the property and the time when the sale took place. This increases the network's robustness against unseen regions or future changes in the market, as it restrains from modelling the overall trend of the market.

To normalise a price (p_i) of the property i under our scheme, we first calculate the average price per m^2 (a^t_m) in town m at year t. Equation 4 illustrates the calculation of a^t_m, where S^t_m is the set of all sales within town m that occurred in year t and f_i is the floor area of a property in m^2:

$$a^t_m = \frac{1}{|S^t_m|}\sum_{i \in S^t_m}\frac{p_i}{f_i} \tag{4}$$

Further, we perform normalisation of the property price by dividing it by f_i and a^t_m. Equation 5 expresses the normalised price (y_i) of a property.

$$y_i = NormalisedPrice(p_i, f_i, a^t_m) = \frac{p_i}{f_i \times a^t_m} \tag{5}$$

This normalised price scheme is inserted in the mean squared error optimization Equation (3).

Equation 6 illustrates how to retrieve the original property price, given the normalised property price:

$$\hat{p}_i = NormalisedPrice'(\hat{y}_i, f_i, a^t_m) = \hat{y}_i \times f_i \times a^t_m \tag{6}$$

We intend to compare our proposed normalization scheme against the logarithm of the price normalisation scheme, due to the matching rationale for using each of them and its prevalence in the literature.

4.4. Evaluation Metrics

To assess the effectiveness of the models in the task of property appraisal, we evaluate our models using mean squared error (MSE), mean absolute error (MAE) and the coefficient of determination (R^2).

MSE (Equation 3) effectively signals the frequency of significant errors, accentuated by the squaring operation, which magnifies the impact of large and infrequent errors. Moreover, it allows clear monitoring of the progress of optimization and allows for early stopping to prevent overfitting.

MAE (Equation 7) enables comparative analysis with the value of MSE. It is insensitive to infrequent large errors but provides a meaningful metric that is simple to interpret and can be simply related to real-world average estimation error.

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{7}$$

R^2 (Equation 8) describes the squared distance of the prediction from the actual values, adjusted by the variance in the actual values. The coefficient of determination can be interpreted as measuring the model's goodness of fit to the true values.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{8}$$

5. Experiments

5.1. Augmenting the Baseline

We first assess the performance of the *skinny* and *augmented* baseline models proposed by Azizi & Rudnytskyi (2022) and Law et al. (2019). We are interested in assessing what performance the baseline models achieve by just using the features (or their proxies) included in the original paper (*skinny* baseline) and their performance when all of the features from the Latvian dataset are provided (*augmented* baseline).

The hyperparameter settings for each of the baselines used in the experiments were the same as those described in the original papers. Azizi & Rudnytskyi (2022) used learning rate = 1.5×10^{-3}, dropout = 0.1 and the AdaMAX optimizer (Kingma & Ba, 2017). Law et al. (2019) used learning rate = 10^{-4}, but no dropout with ADAM optimizer (Kingma & Ba, 2017). Neither of the studies mentioned L2 regularization.

The results show that the *augmented* baselines (Table 4 indices 2, 5) performed much better on the test set compared to the *skinny* baselines (Table 4 indices 1, 4). The *augmented* baselines achieved a 15 and 36 percentage points greater MSE for Azizi & Rudnytskyi (2022) and Law et al. (2019) networks when evaluated on the random city test set. When assessing the models' performance on the unseen town test set, Law et al.'s (2019) model improved by a small margin, whilst Azizi & Rudnytskyi (2022) became significantly worse (20 percentage points greater MSE). This can be partly explained by the effect of integrating POI and transport maps, which are described more in detail in Section 5.3.

It can be concluded that the increase in input feature set size increased the model's ability to generate predictions of greater quality. It has to be noted that some of the results of the R^2 were adjusted (denoted with a '*') due to model outputs yielding infinities as a result of logarithm price exponentiation. In those cases, the prediction was replaced with a prediction of 1, which is favourable for the model.

In conclusion, the inclusion of all available features from our dataset was beneficial to the baseline models. Therefore, all further experiments will be performed on the 'augmented' set of features to ensure consistency between models and maximise the performance of each model type.

5.2. Geo-temporal objective function on the Baseline

Further experiments investigate the effect of the proposed geo-temporal objective function introduced in Section 4.3.3. To test the effectiveness of this objective function, we have trained and compared both baseline models when modelling the logarithm of price (Table 4, indices 2 and 5) and modelling the geo-temporally normalised price (Table 4, indices 3 and 6) using MSE. We reuse the same hyperparameter settings from the previous experiment (see Section 5.1).

The outcomes of the experimentation exhibit the benefits of the geo-temporal objective function as they have allowed previously underperforming models to fit the diverse and difficult dataset of the task given. For both baselines, the use of geo-temporally normalised prices has reduced the MSE loss threefold on the random property test set and fourfold on the unseen town test set compared to the log price modelling. This is consistent with our expectations, especially when it comes to the unseen town test set performance, as the model was capable of performing well by obtaining R^2 of over 65% while being unaware of the market dynamics. Interestingly, the geo-temporally normalised price was better at modelling the sample in the unseen town test set than the random town test set by a margin of 3 and 10 percentage points in the case of Azizi & Rudnytskyi (2022) and Law et al. (2019) respectively. This can explained by the fact that the samples in the unseen town test are on average cheaper than the samples in the random test set. Therefore this discrepancy is consistent with our expectation of the geo-temporarily normalised price, which puts greater emphasis on cheaper properties.

Random Test Set

	A	Features	Obj.	MSE	MAE	R^2
1	**1**	$T_S + S_{80}$	LP	0.8077	0.7518	-0.1997
2		ALL		0.6575	0.6739	-0.2012
3		ALL	GT	0.2669	0.4051	0.6245
4	**2**	$T_S + S_{300}$ + POI	LP	1.0062	0.7826	-0.7892
5		ALL		0.6422	0.5976	-0.1651*
6		ALL	GT	0.2694	0.4084	0.5607

Unseen Town Test Set

	A	Features	Obj.	MSE	MAE	R^2
1	**1**	$T_S + S_{80}$	LP	1.4833	1.0114	-0.3860
2		ALL		1.7025	1.1148	-0.4635
3		ALL	GT	0.2583	0.4074	0.6550
4	**2**	$T_S + S_{300}$ + POI	LP	1.0042	0.8012	-0.7268
5		ALL		0.9651	0.8118	0.0136*
6		ALL	GT	0.2276	0.3789	0.6645

Table 4. Test performance of Azizi & Rudnytskyi (2022) (Author 1) and Law et al. (2019) (Author 2) baseline models, assessed by mean squared error (MSE), mean absolute error (MAE) and R^2 score. For feature abbreviations, refer to Tables 1 and 2. LP stands for logarithm price, whereas GT stands for geo-temporal normalised price. ALL = T + S_{80} + S_{300} + POI + TM

5.3. Benefits of including POI and transport map data

This section discusses the effects of POI and transport maps being included as input features to perform real estate appraisal. To assess the effects of additional feature integration, we train the *Fusable Image Model* introduced in Section 4.2. We analyse two versions of this model - one trained with 2 satellite images (S_{80}, S_{300}) and the other trained with 2 satellite images, POI and transport map (S_{80}, S_{300}, POI, TM). For the models' results refer to Table 5, indices 2 and 3.

The experiments have yielded interesting contradictory results. There was a minor improvement of 1.3% in terms of the random test set R^2, whilst there was a minor decrease of 1.7% in performance in terms of unseen town test R^2 metric. We explain the performance decrease on the unseen town test set with the lack of information in the respective POI and transport maps, as the investigated cities lack many amenities and transportation links that other cities in the random dataset possess. This inherently limits the model's ability to extract valid information about those places, leading to decreased performance.

We are unable to draw any concrete conclusion on the benefits of including POI and transport map data in the model. However, based on the slightly biased nature of or unseen town test set, we decided to include this data in the final model of this study, based purely on the increase in performance in terms of the less biased random test set.

Random Test Set

	Features	MSE	MAE	R^2
1	T	0.2040	0.3285	0.7190
2	S_{80} + S_{300}	0.2636	0.3917	0.6442
3	S_{80} + S_{300} + POI + TM	0.2590	0.3933	0.6530
4	T+S_{80} + S_{300} + POI + TM	0.2220	0.3519	0.6933

Unseen Town Test Set

	Features	MSE	MAE	R^2
1	T	0.2115	0.3563	0.7181
2	S_{80} + S_{300}	0.2452	0.3900	0.6214
3	S_{80} + S_{300} + POI + TM	0.2664	0.4036	0.6107
4	T+S_{80} + S_{300} + POI + TM	0.2099	0.3564	0.7287

Table 5. Test performance of the proposed model, assessed by mean squared error (MSE), mean absolute error (MAE) and R^2 score. For feature abbreviations, refer to Tables 1 and 2.

5.4. Multimodal model with geo-temporal objective

Finally, we assess the performance of the deep multi-modal architecture introduced in Section 4.2, on the task of minimising MSE loss on the geo-temporal objective.

The training of the deep multi-modal network involved a pre-training step of the *Fusable Tabular Model* (Table 5, index 1) and *Fusable Image Model* (Table 5, index 3). After, the weights of those models were transferred to the new model in the manner exhibited in Figure 2. After the weight transfer, the model was trained for 15 epochs and early stopped at epoch 13 when it achieved the highest validation loss. The hyperparameter choice for the proposed architecture was outlined in 4.2.

The proposed model has significantly outperformed all other approaches in terms of the unseen town test, by achieving R^2 of 0.7287, whilst even the best baseline using geo-temporally normalised price had achieved 0.6645. In terms of the random test set the *Fusion Model* achieved R^2 of 0.6933, whilst the best baseline only achieved 0.6245. Interestingly, the *Fusable Tabular Model* was able to outperform the *Fusion Model* when tested on a random test set by 3.7%.

Based on the fact that the random test set performance of the *Fusion Model* is still very good (0.6933 R^2), whilst performing much better on the unseen town set (0.7287 R^2), we hypothesise that this model is likely to generalize better than *Fusable Tabular Model*. Nevertheless, we argue that *Fusable Tabular Model* is a great choice for the case with limited computation resources or no access to image data.

6. Future work and limitations

To assess the proposed model's ability to generalise across different property markets, we could evaluate its performance on property markets of the neighbouring countries. Additionally, when tackling those markets, the presented model could be used as a pre-training step to a model that is fine-tuned to the specific country market. Additionally, due to their prevalence in real estate appraisal literature, classical machine learning approaches could be employed in conjunction with the proposed model, where the neural model is used as a feature extractor.

The dataset acquired for this task is not of the highest quality, as many unlikely values were observed and tax fraud is suspected to have occurred. Therefore, additional work in dataset cleaning and validity checking is required. A correct and consistent dataset is likely to lead to large model performance improvement. Lastly, the temporal aspect of the proposed objective function has not been evaluated to a satisfactory degree. Therefore a separate test set 'unseen year(s)', would evaluate the models' ability to generalise over different time periods.

7. Conclusions

In this paper, we have tackled deep-learning-based apartment value prediction using multi-modal attributes. As the majority of the real estate appraisal literature has previously focused on using relatively shallow architectures, we introduce a deeper network, encompassing DenseNet (Huang et al., 2017) CNN architecture, transfer learning techniques and increased added L2 and batch normalisation regularisation methods. Our proposed model achieves a 0.7287 R^2 score on the unseen town test set, outperforming baseline models from the deep neural network multi-modal real estate appraisal literature (Azizi & Rudnytskyi, 2022; Law et al., 2019) by 0.1042 and 0.1680, respectively.

Additionally, we devised a novel geo-temporal normalisation scheme that abstracts away from the location and time of sale of the property, creating a general model that abstains from modelling the market dynamics. The viability of the scheme was assessed by comparing the model performance with the logarithm normalisation scheme on the baseline models. Both Azizi & Rudnytskyi (2022) and Law et al. (2019) baseline models demonstrated a significant performance improvement when using our geo-temporally normalised objective function.

Lastly, we analysed the benefits of using point-of-interest maps and transport maps as additional input to multi-modal property appraisal models. This analysis has shown mixed results, disallowing us to concretely conclude about their importance.

All models were trained and evaluated on a new dataset, based on a publicly available real estate market database created by the State Land Service of Latvia and self-collected satellite and map imagery.

References

Azizi, Ilia and Rudnytskyi, Iegor. Improving real estate rental estimations with visual data. *Big Data and Cognitive Computing*, 6(3), 2022. ISSN 2504-2289. doi: 10.3390/bdcc6030096.

Baum, Andrew, Graham, Luke, and Xiong, Qizhou. The future of automated real estate valuations (avms), October 2021.

Bency, Archith J., Rallapalli, Swati, Ganti, Raghu K., Srivatsa, Mudhakar, and Manjunath, B. S. Beyond spatial auto-regressive models: Predicting housing prices with satellite imagery. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 320–329, 2017. doi: 10.1109/WACV.2017.42.

Bin, Junchi, Gardiner, Bryan, Liu, Zheng, and Li, Eric. Attention-based multi-modal fusion for improved real estate appraisal: a case study in los angeles. *Multimedia Tools and Applications*, 78(22):31163–31184, 2019. ISSN 1573-7721. doi: 10.1007/s11042-019-07895-5.

Chen, Meixu, Liu, Yunzhe, Arribas-Bel, Dani, and Singleton, Alex. Assessing the value of user-generated images of urban surroundings for house price estimation. *Landscape and Urban Planning*, 226:104486, 2022. ISSN 0169-2046. doi: https://doi.org/10.1016/j.landurbplan.2022.104486.

Huang, Gao, Liu, Zhuang, Van Der Maaten, Laurens, and Weinberger, Kilian Q. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017. doi: 10.1109/CVPR.2017.243.

Ioffe, Sergey and Szegedy, Christian. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015.

Jiang, Chao, Jiang, Canchen, Chen, Dongwei, and Hu, Fei. Densely connected neural networks for nonlinear regression. *Entropy*, 24(7), 2022. ISSN 1099-4300. doi: 10.3390/e24070876.

Kingma, Diederik P. and Ba, Jimmy. Adam: A method for stochastic optimization, 2017.

Law, Stephen, Paige, Brooks, and Russell, Chris. Take a look around: Using street view and satellite images to estimate house prices. *ACM Transactions on Intelligent Systems and Technology*, 10(5):1–19, September 2019. ISSN 2157-6912. doi: 10.1145/3342240.

Ozer, Faruk and Okan Sakar, C. An automated cryptocurrency trading system based on the detection of unusual price movements with a time-series clustering-based approach. *Expert Systems with Applications*, 200:117017, 2022. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2022.117017.

Simonyan, Karen and Zisserman, Andrew. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Srivastava, Nitish, Hinton, Geoffrey, Krizhevsky, Alex, Sutskever, Ilya, and Salakhutdinov, Ruslan. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56): 1929–1958, 2014.

Szegedy, Christian, Vanhoucke, Vincent, Ioffe, Sergey, Shlens, Jon, and Wojna, Zbigniew. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016. doi: 10.1109/CVPR.2016.308.

Vivian W. Y. Tam, Ivan W. H. Fung, Jing Wang and Ma, Mingxue. Effects of locations, structures and neighbourhoods to housing price: an empirical study in shanghai, china. *International Journal of Construction Management*, 22(7):1288–1307, 2022. doi: 10.1080/15623599.2019.1695097.

Yang, Linchuan, Chau, K.W., Szeto, W.Y., Cui, Xu, and Wang, Xu. Accessibility to transit, by transit, and property prices: Spatially varying relationships. *Transportation Research Part D: Transport and Environment*, 85, 2020. doi: 10.1016/j.trd.2020.102387.