

Received February 27, 2021, accepted March 30, 2021, date of publication April 6, 2021, date of current version April 15, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3071306

Deep Learning Model for House Price Prediction Using Heterogeneous Data Analysis Along With Joint Self-Attention Mechanism

PEI-YING WANG¹, CHIAO-TING CHEN¹, JAIN-WUN SU¹, TING-YUN WANG¹, AND SZU-HAO HUANG¹, (Member, IEEE)

¹Institute of Information Management, National Chiao Tung University, Hsinchu 30010, Taiwan

²Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan

³Department of Information Management and Finance, National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan

Corresponding author: Szu-Hao Huang (szuhaohuang@nctu.edu.tw)

This work was supported in part by the Ministry of Science and Technology, Taiwan, under Contract MOST 110-2622-8-009 -014 -TM1, Contract MOST 109-2221-E-009 -139, Contract MOST 109-2622-E-009 -002 -CC2, and Contract MOST 109-2218-E-009 -015; and in part by the Financial Technology (FinTech) Innovation Research Center, National Yang Ming Chiao Tung University.


ABSTRACT House price prediction is a popular topic, and research teams are increasingly performing related studies by using deep learning or machine learning models. However, because some studies have not considered comprehensive information that affects house prices, prediction results are not always sufficiently precise. Therefore, we propose an end to end joint self-attention model for house prediction. In this model, we import data on public facilities such as parks, schools, and mass rapid transit stations to represent the availability of amenities, and we use satellite maps to analyze the environment surrounding houses. We adopt attention mechanisms, which are widely used in image, speech, and translation tasks, to identify crucial features that are considered by prospective house buyers. The model can automatically assign weights when given transaction data. Our proposed model differs from self-attention models because it considers the interaction between two different features to learn the complicated relationship between features in order to increase prediction precision. We conduct experiments to demonstrate the performance of the model. Experimental data include actual selling prices in real estate transaction data for the period from 2017 to 2018, public facility data acquired from the Taipei and New Taipei governments, and satellite maps crawled using the Google Maps application programming interface. We utilize these datasets to train our proposed and compare its performance with that of other machine learning-based models such as Extreme Gradient Boosting and Light Gradient Boosted Machine, deep learning, and several attention models. The experimental results indicate that the proposed model achieves a low prediction error and outperforms the other models. To the best of our knowledge, we are the first research to incorporate attention mechanism and STN network to conduct house price prediction.

INDEX TERMS House price prediction, heterogeneous data, Google satellite map, spatial transformer network, joint self-attention mechanism.

I. INTRODUCTION

House price prediction is currently a hot topic. The purpose of house price prediction is to provide a basis for pricing between buyers and sellers. By viewing transaction records, buyers can understand whether they have received a fair price for a house, and sellers can evaluate the price at

which they can sell a house along a specific road section. House price prediction is also used in financial technology applications, which require a reasonable evaluation system for mortgage calculation and house auctions. Studies on house price prediction have been conducted in the United States [1]–[3] and Europe [4], [5], and most of these studies have adopted machine learning–based methods to predict real estate prices. Related research has also been established even in developing countries such as Brazil [6]. Accordingly, one

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Afzal .

of the objectives of the present study is to use actual selling prices in real estate transaction data to perform predictions. After reviewing research from these countries, we identify several factors that affect house prices and categorize such factors into two groups. The first group comprises house conditions, such as age, building material, building area, and parking space; the second group comprises conditions of the surrounding environment, including public facilities, transportation facilities, and educational institutions. Increasing numbers of studies on house price estimation have adopted machine learning and deep learning methods; researchers have mostly been concerned with the accuracy of forecasting results. However, some studies have considered only house attributes, and they have not accounted for information on surrounding environments. Therefore, using the Google Maps application programming interface (API), we crawl satellite maps of houses and surrounding public facilities as heterogeneous input data for the prediction model. We use house transaction data, the satellite maps, and public facility data to experimentally test our hypothesis. To process the satellite maps, we adopt a spatial transformer network (STN) [7] to extract image features. An STN has rotation-invariant properties, meaning that image extraction features do not change markedly when an image is rotated by a specific angle.

Different groups of buyers may focus on different house attributes. For example, nuclear families may focus on parks and nearby schools for their children. Therefore, if a house's crucial features identified through an attention mechanism match the needs of a prospective buyer, then the house can be recommended to the buyer. Accordingly, our final research goal is to improve prediction accuracy and investigate features influencing pricing results. We employ an attention mechanism and import heterogeneous data into our developed attention model. These data and the introduced methods not only improve house price prediction performance but also reveal factors influencing house prices. In addition to achieving accurate house price prediction, we identify the features contributing to house prices. Therefore, we develop an attention-based model that can learn the interaction between features and summarize crucial attributes for house buyers. Additionally, we design novel methods to effectively extract features from heterogeneous data to help improve model performance.

The main contributions of our framework can be summarized as follows:

- We import heterogeneous data comprising Google satellite maps and public facilities to prove that data on house transactions alone are inadequate for achieving high prediction performance. Therefore, adding heterogeneous data can improve data diversity and prediction performance.
- We combine STN network to obtain image features, which include all information on elements around a house in the satellite map. Because of its main

characteristic of rotation invariance, the STN can provide image features that do not change markedly.

- We propose a joint self-attention mechanism that not only includes an attention mechanism but also considers two-hop relationships between different attributes to identify their implicit relationships and improve model training performance.
- To the best of our knowledge, we are the first research to incorporate attention mechanism and STN network to conduct house price prediction.

II. RELATED WORKS

This chapter first introduces current studies on house price prediction, which are categorized into those using deep learning and machine learning methods. Subsequently, we describe the role of attention mechanisms in various applications. Finally, we provide an overview of relevant studies to summarize the present work.

A. HOUSE PRICE PREDICTION

Some studies on house price prediction have adopted machine learning methods such as support vector machine (SVM) and extreme gradient boosting (XGBoost) algorithms, and other studies have utilized deep learning methods such as long short-term memory (LSTM) networks or convolutional neural networks (CNNs). In addition, some studies have not only considered house attributes but also accounted for heterogeneous data, including street-view or satellite maps. The following paragraphs briefly introduce these related studies.

Mu *et al.* [2] applied SVM and least squares SVM methods to perform house price prediction and demonstrated that both methods outperformed partial least squares. Peng *et al.* [8] adopted an XGboost algorithm to predict secondhand house data in Chengdu, China, and they demonstrated that XGboost can make a model more robust and prevent overfitting compared with a decision tree and linear regression. Madhuri *et al.* [9] employed multiple regression techniques such as least absolute shrinkage and selection operator, gradient boosting, and adaptive boosting for house price forecasting. Their proposed model could help sellers estimate costs of sales and provide exact information for buyers. Bork and Møller [3] predicted real estate prices in all 50 states in the United States and utilized a dynamic model selection mechanism for forecasting. Phan [10] proposed a novel model that integrates machine learning methods, including SVM and stepwise methods, to reveal beneficial information from past transaction data in Melbourne, Australia.

Temur *et al.* [11] combined an autoregressive integrated moving average model and an LSTM network to develop a novel model for house price prediction in Turkey and other countries. They also adopted mean absolute percentage error (MAPE) and mean squared error as evaluation metrics. Compared with other models, their hybrid model was proven to have a lower error rate and accurate prediction results. Yu *et al.* [12] developed a deep learning-based model combining an LSTM network and a CNN to forecast

secondhand house prices in Beijing, China; to demonstrate the effectiveness of the proposed model, they compared the performance of this model with that of an autoregressive moving average model. Ge [13] proposed a novel framework learnable graph convolutional network combined with a Graph CNN and LSTM network to capture temporal–spatial features, economic factors, and community characteristics in order to improve house price forecasting. Afonso *et al.* [6] proposed a model combining a random forest and a recurrent neural network (RNN) and applied it to a Brazilian house dataset.

Law *et al.* [14] used features obtained from street view and satellite images to capture urban characteristics in London and improve house price prediction. Fu *et al.* [15] integrated house transaction data obtained from a website with geolocation data and point of interest data and developed a novel architecture for open access dataset–based hedonic price modeling to analyze the real estate markets in Shanghai and Beijing, China. You *et al.* [16] employed RNNs for house price prediction by adopting the visual features of corresponding houses; their experimental results indicated that the inclusion of visual features in their model reduced the MAPE and mean absolute error of the model compared with those of other state-of-the-art (SOTA) models do. Poursaeed *et al.* [17] hypothesized that the internal and external appearances of a house are key factors affecting house price. Therefore, they developed a novel model including images of internal and external house features and conducted experiments on the Zillow, Redfin, and Trulia real estate databases; the experimental results indicated that their model outperformed other prediction models.

B. ATTENTION MECHANISMS

To identify major factors influencing house prices, we adopt attention mechanisms. Attention mechanisms are widely used in many fields because of their ability to distinguish features, such as machine translations, image captions, and speech recognition. Such mechanisms can enhance crucial parts of a system by assigning higher weights to more influential features, and they can block irrelevant parts by assigning lower weights to less relevant features in the model. The following paragraphs introduce some relevant studies that have used attention mechanisms.

A previous study on attention-based machine translation [18] used an attention mechanism to improve translation accuracy. The study proposed two attention mechanisms for use with source words and a subset of source words, and this architecture was applied to translation between German and English. Vaswani *et al.* [19] adopted a self-attention mechanism to develop transformer architecture for two translation tasks, namely English to German and French to English. Firat *et al.* [20] developed a multilingual machine translation model that could translate a single language into multiple languages. The model has a shared attention mechanism, demonstrating that the proposed model can exhibit high performance in translation quality. Huang *et al.* [21] developed

a neural machine translation architecture that exploits visual features with text features for translation, and they adopted an LSTM network to assist in generating a representation of an attention-based encoder.

Regarding speech recognition, [22] extended the concept of attention to Texas Instruments/Massachusetts Institute of Technology phoneme recognition, meaning that each term in a sentence can be weighted differently to increase the flexibility of neural network model learning. Chorowski *et al.* [23] developed a novel model for sequence modeling that adopts an attention mechanism, and they utilized an RNN to reveal the relationship between an input sequence and output. The experimental results demonstrated that the attention mechanism could reduce the operating time and perform similarly to a hidden Markov Model (HMM)-free RNN-based method. Moritz *et al.* [24] introduced a speech recognition architecture called triggered attention, which combines an attention mechanism and a connectionist temporal classification (CTC) method; they tested it on three datasets in different languages. Watanabe *et al.* [25] proposed a hybrid model combining CTC and an attention mechanism and imported a multiobjective learning model to increase robustness and convergence. The experimental results indicated that their proposed model outperformed fully connected networks and HMM models.

You *et al.* [26] introduced a novel algorithm combining top-down and bottom-up approaches through a semantic attention model. They also experimented with two classic benchmarks, namely Flickr30K and Microsoft Common Objects in Context, by using the proposed algorithm. Zhou *et al.* [27] proposed an attention model combining a convolution layer and a convolutional LSTM network to extract spatiotemporal features. The proposed model captures pickup and dropoff interactions and predicts passenger pickup and dropoff demand for taxi services. Gao *et al.* [28] proposed a novel framework composed of an attention mechanism and an LSTM network to capture salient objects from short videos and the relationships between words and videos. Yang *et al.* [29] developed multiple-layer stacked attention networks (SANs) for image question answering. Experimental results demonstrated that the SANs could outperform other SOTA approaches and visualize how the attention layer reaches the answer to a question.

In addition, multimodal data fusion across various domains has attracted considerable scholarly attention in recent years [30]. Many attention mechanisms use heterogeneous features for the incorporation of more complete knowledge. Hori *et al.* [31] reported that their multimodal attention model, constructed by using a combination of audio, image, and motion features, achieved excellent performance. Xu *et al.* [32] presented a framework involving the learning of Multimodal Attention Long-Short Term Memory Networks (MA-LSTM) to enhance the automatic generation of video captions. Yue *et al.* [33] proposed using the HE-CLSTM, a privacy-preserving, computer-aided diagnosis algorithm, to analyze encrypted time-series medical images.

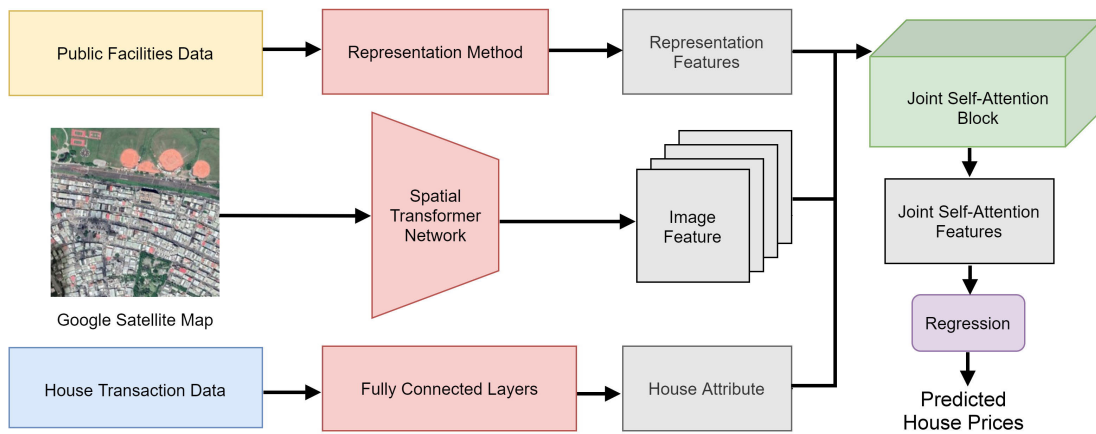


FIGURE 1. System framework overview.

Woo *et al.* [34] developed the CBAM, which can be effectively integrated into any CNN architecture. The results indicated that the CBAM achieved better performance on classification tasks than did the other models examined.

After reviewing relevant studies, we crawl satellite maps to obtain the terrain surrounding houses; these map data can reveal green space, coasts, and schools near houses. In addition to the image dataset, public facility data that can represent the amenities near houses are retrieved as part of the input data. A house transaction dataset, public facility dataset, and satellite map dataset are used as our experimental data. Inspired by the concept of attention, we adopt an attention mechanism to automatically assign weights to different features; this can thus avoid the use of only the sample model for house price prediction, considering the fact that complex relationships exist among factors in the housing market.

III. PROPOSED METHODOLOGIES

This section summarizes our house price prediction system, which comprises a joint self-attention mechanism and some special feature representation methods. Subsequently, we briefly introduce the components of our system. Our system can be separated into several components, and we provide the core concept of each component's function.

A. SYSTEM OVERVIEW

This section provides an overview of our house price prediction framework. Moreover, each of the components of our system is briefly introduced in the subsequent sections. Fig. 1 presents an overview of our system; specifically, the figure illustrates the overall concept of our system. Our proposed system can be separated into three components: a public facility data feature representation method, an image feature extractor adopting an STN, and a joint self-attention mechanism for identifying major factors that affect house prices.

B. PUBLIC FACILITIES DATA REPRESENTATION METHODS

When buying a house, people not only focus on its internal features, such as its age, pattern, building material, and building area, but also consider its surrounding environment. However, real estate transaction data provided by the Ministry of the Interior of Taiwan record only the aforementioned internal features. These features do not comprehensively constitute information on a house. Therefore, we crawl data on public facilities surrounding houses to gather complete information.

We then identify public facilities that buyers notice. Different buyers focus on different facilities. Office workers might prefer to have a convenience store nearby for food, a transportation system near their homes for their commute to work, and parks for exercise or walks after work. Older adults may focus on hospitals near their homes to reduce their distance to doctors. Considering these conditions, we crawl related information from the New Taipei City and Taipei governments to obtain our model input data. We acquired the open access data from platforms maintained by the Taipei City (<https://data.taipei/#/>) and New Taipei City (<https://data.ntpc.gov.tw/>) governments; access to data was easy and convenient. The Taipei City platform provides open access data on 22 categories, including transportation, economics, medicine, and agriculture. The New Taipei City platform also provides open access data on 22 categories, with the most popular categories including statistics, transportation, information, finance and taxation, land administration, and medicine.

We transform house and public facility addresses into latitudes and longitudes and then calculate the distance between a house and each public facility. Generally, public facilities that are only within 1 km of a house are counted; however, through this method, we cannot determine where these amenities are distributed around the house. To address this, we change the representation method (Fig. 2) to use four quadrants and three distances; a total of 12 columns are used to calculate the number of public facilities in different ranges at different angles. Concentrating many convenience

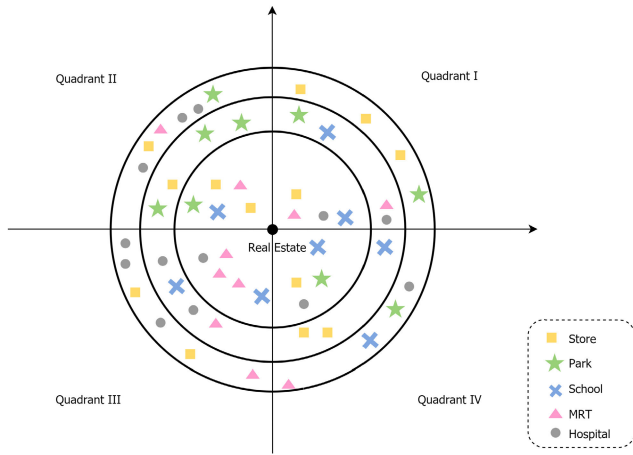


FIGURE 2. Proposed representation method for extracting public facility data features.

stores or mass rapid transit (MRT) stations in a certain direction may be unfavorable because amenities around a house that is located on the edge of the city would be inadequate. Therefore, to help identify a house that is favorably located in terms of surrounding amenities, this method considers facility distribution, distance, and quantity to describe the characteristics of various facilities surrounding a house. Through this method, we can record the number of parks, MRT stations, hospitals, convenience stores, and schools surrounding a house within specific distances and quadrants to fully illustrate the situation of the surrounding environment.

C. IMAGE FEATURE EXTRACTOR

This section introduces how we extract satellite map features by adopting an STN [7]. We crawl satellite maps when given a specific address by using the Google Map API. Most studies have used CNNs to extract image features. Conventional CNNs use convolutional and pooling layers to achieve rotation invariance to an extent; however, artificially set transformation rules are required for this type of rotation invariance, and such rules cannot produce complete rotation invariance. As Fig. 3 presents, the two photos are actually of the same location; a baseball field is visible in front of the house in Fig. 3a. After the image is rotated (Fig. 3b), the baseball field appears on the right-hand side; consequently, features extracted through the CNN will be different from those obtained for the unrotated image. In reality, when buying a house, buyers consider relationships in all directions to be consistent. Therefore, our model requires a rotation-invariant CNN that would not affect model judgment even when images are rotated. We thus adopt an STN to extract image features, and the STN architecture comprises three modules (Fig. 4). Each module is composed of a localization network, a grid generator, and an image sampler, and they are individually introduced as follows:

The localization network is used to transform and learn the feature map. The input image is $I \in R^{h \times w \times c}$, where w represents width, h represents height, and c represents channels.



FIGURE 3. Example of a rotated satellite map.

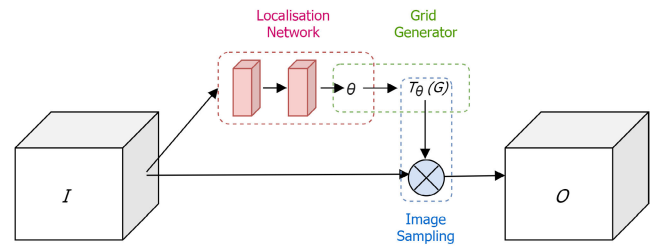


FIGURE 4. STN architecture [7].

The output image is $O \in R^{h' \times w' \times c'}$. The input image is subjected to several iterations of processing through layers such as convolutional and max pooling layers, and the feature map is transformed by the localization network to obtain θ , where $\theta = f_{ln}(I)$ and θ represents the parameter P_θ , which is used in the subsequent part.

The grid generator uses the output θ of the localization layer to transform the feature map; it then performs a spatial transformation to correspond a spatial location in the output feature map to that in the input feature map. The mapping between the input and output feature maps is conducted using T_θ through an affine transformation Aff_θ expressed in (1).

$$\begin{pmatrix} x_i^m \\ y_i^m \end{pmatrix} = T_\theta(Grid_i) = Aff_\theta \begin{pmatrix} x_i^n \\ y_i^n \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^n \\ y_i^n \\ 1 \end{pmatrix} \quad (1)$$

After these two steps, each pixel of the output feature map is corresponded to a position in the input feature map through spatial transformation; however, because the coordinates are often not integers, the partial derivative of the feature score cannot be calculated. Therefore, bilinear interpolation is adopted for mapping, as expressed in (2), where O_i^c represents the output value of pixel i at location (x_i^t, y_i^t) in channel c and I_{ab}^c represents the value at point (a, b, c) . A pixel position (coordinates) can be calculated using the max function; scores corresponding to four points around (x_i^m, y_i^m) can be selected to calculate a final score that represents the pixel position.

$$O_i^c = \sum_a^h \sum_b^w I_{ab}^c \max(0, 1 - |x_i^m - b|) \max(0, 1 - |y_i^m - a|) \quad (2)$$

When the STN is applied to Google satellite maps, it can learn the same features, regardless of whether an image is rotated. Therefore, it can be easily embedded into the current CNN model to achieve end-to-end training.

D. JOINT SELF-ATTENTION MECHANISM

We adopt a self-attention mechanism [19]. Self-attention differs from conventional attention mechanisms because it is adopted on the source side and target side and is performed only on data related to the input of the source side or input of the target side to capture relationships between words on the source or target side. Subsequently, attention features obtained from the source side are integrated with those retrieved from the target side to capture the relationship between words on the source and target sides. Therefore, self-attention mechanisms outperform conventional attention mechanisms. One of the main reasons is that conventional attention mechanisms often ignore dependencies between words in source or target sentences. However, self-attention mechanisms can obtain not only dependencies between words on the source and target sides but also dependencies between words in the source or target. Because of these advantages, self-attention mechanisms are used in transformer modules to solve translation tasks.

In a self-attention mechanism for a sentence, for example, each word X can be transformed into three elements, namely $Query$, Key , and $Value$, using (3). W_Q , W_K , and W_V represent three different transformation matrices of X . The self-attention weights can be calculated using (4). The term $(\frac{Query Key^T}{\sqrt{d_k}})$ is a matching function that expresses the matching level between the query and the key, and d_k represents the dimension of Key .

- $Query$ represents the query, which means that each word can be a question that can be asked;
- Key represents the key that is matched with each query; and
- $Value$ represents the value that extracts the most crucial information from each word.

$$\begin{aligned} Query &= W_Q X \\ Key &= W_K X \\ Value &= W_V X \end{aligned} \tag{3}$$

$$(\text{Attentive Weights} = \text{softmax} \left(\frac{Query Key^T}{\sqrt{d_k}} \right) Value) \tag{4}$$

Inspired by the concept of self-attention, we develop a novel framework (Fig. 5). In this framework, a self-attention mechanism is used to transform the study problem. $HouseAttr.$ is transformed into a query, and $attribute$ is transformed into a key and value simultaneously. Our purpose is to identify attributes that affect house prices. However, if we adopt this setting, we can identify only the most crucial types of attributes, meaning that we will miss the interactions between different attributes, namely the two-hop relationships. Therefore, we improve the key and value function to contain not only one-attribute relationships but

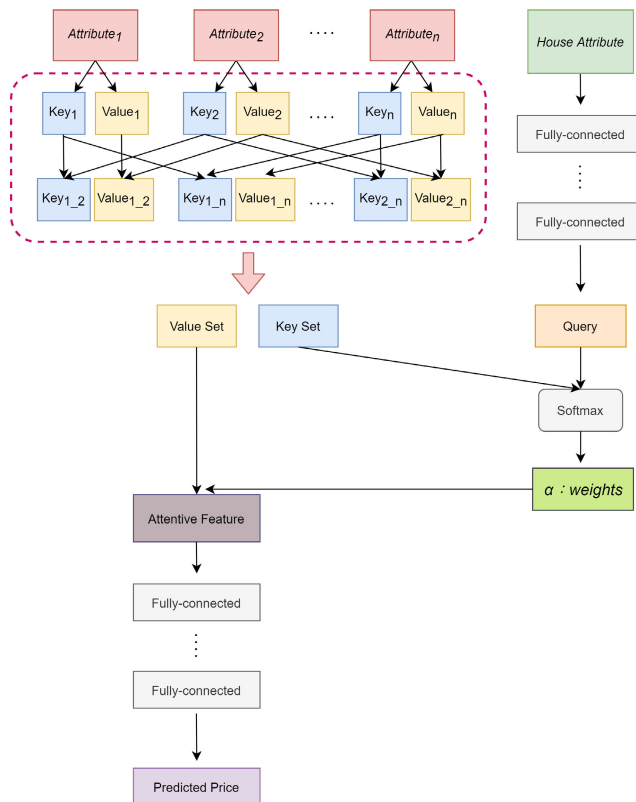


FIGURE 5. House price prediction based on a joint self-attention mechanism.

also two-attribute interactions. We adopt this novel method to generate the key and value because house pattern and location may be crucial individual factors influencing house prices; these two features may also constitute a major reason for purchasing a house. Therefore, we consider not only single attributes but also the interaction between several attributes to improve model performance. Under the joint self-attention mechanism, the house transaction data were first transformed into queries through several fully connected layers. All the attributes of these data were transformed into keys and values. We then obtained the information common to key–value pairs by calculating their relationship and incorporating two of them into a value and a key set. Next, the key set was used to produce attention weights by using a softmax layer. Finally, the attention weights were multiplied by the corresponding values and went through several fully connected layers to generate a predicted price. Algorithm 1 presents the joint self-attention mechanism.

E. PREDICTION MODEL

According to the data limitations, road sections can be separated into four addresses, and a gated neural network can then be adopted to produce the respective weights and identify possible house locations. Four datasets are input into the gated mechanism to generate weights to represent possible addresses that a house may have. The weights are generated through the process illustrated in Fig. 6. A gated neural

Algorithm 1 Joint Self-Attention Mechanism

Input: House transaction data *House Attr.*, all attributes attribute

Output: Embedding of all attributes *emb*

- 1: Transform *House Attr.* into *Query*
- 2: **for** $n = 1, 2, \dots, N$ **do**
- 3: Transform attribute into Key_n ;
- 4: Transform attribute into $Value_n$;
- 5: **end for**
- 6: Calculate $Key_{inter.}$ interaction between two different keys;
- 7: Calculate $Value_{inter.}$ interaction between two different values;
- 8: Append $Key_{inter.}$ with all Key_n into Key ;
- 9: Append $Value_{inter.}$ with all $Value_n$ into $Value$;
- 10: Obtain the attention weight between *Query* and all Key by using the scaled-dot product;
- 11: Obtain the joint self-attention weight *attentive weight* of all attributes by multiplying the weights by the corresponding $Value$;
- 12: **return** *attentive weight*;

network is a basic attention model that uses simple methods to calculate attention weights when given various types of house transaction data. In this case, because the transaction data did not record the exact addresses—for example, an address may appear as “No.1–No. 30 Kangding Rd., Wanhua Dist., Taipei City, 10843, Taiwan (R.O.C.)”—we used a Google API to crawl the latitude–longitude coordinates of the addresses through two methods, one of which involved dividing the sections into four addresses, as introduced in p.8. Next, the gated neural network was used to generate weights for these four addresses so that their importance could be determined. In essence, the four addresses first went through four independent networks and output four predicted prices. The gated neural network then output four weights that represented the respective importance of the four addresses. $Network_1, Network_2, Network_3, Network_4$ represent the four respective datasets of the four addresses. These data are processed through several fully connected layers to retrieve their hidden information and then processed using the softmax function to obtain α , which is multiplied by the predicted prices of the different input data in (5). The gated model is established with fully connected neural networks, and attention weights α are derived using the fully connected network θ and softmax activation function σ . The final house price $Price$ is calculated by multiplying the price in each network by the respective attention weights as expressed in (6).

$$\alpha = \frac{\exp(\sigma(\theta(\text{Input Data})))}{\sum \exp(\sigma(\theta(\text{Input Data})))} \quad (5)$$

$$Price = \sum_{i=1}^4 Price_i * \alpha_i \quad (6)$$

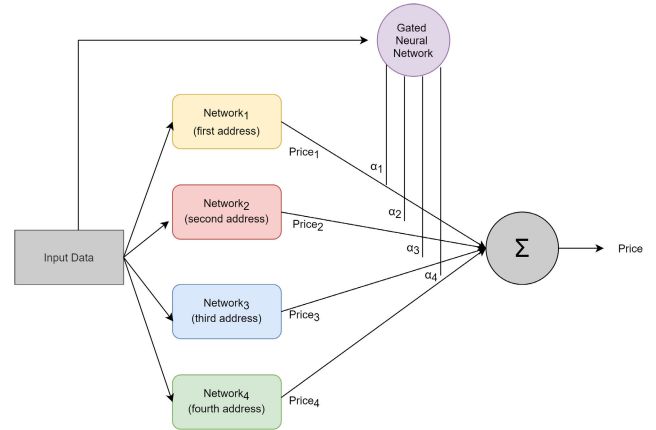


FIGURE 6. Weights generated by a gated neural network for house price prediction.

In sum, this study develops a joint self-attention-based model that can be separated into three parts. First, the public facility data representation method can transform data on public facilities such as schools, hospitals, MRT stations, parks, and convenience stores into useful information to supplement the house transaction data. Second, the STN extract crucial features from satellite maps by capitalizing on its rotation invariance. Third, the joint self-attention mechanism applies queries, keys, and values to derive the weight of each attribute; moreover, two-hop information is considered in the weight derivation process to identify crucial factors that influence house buyers.

IV. EXPERIMENT

We conduct four experiments to evaluate the usefulness of our proposed model. In this section, we describe these experiments in detail and introduce the evaluation metrics used to assess the performance of all approaches in the experiments. We also describe the experimental settings and results. In the first experiment, a basic deep-learning model is developed, and heterogeneous data comprising public facilities and satellite maps are imported to verify whether the use of such data can improve prediction accuracy. In the second experiment, several attention mechanisms are adopted for comparison with the results of the joint self-attention mechanism. In the third experiment, we apply our proposed model to analyze house prices in Taichung and Kaohsiung, Taiwan. In the final experiment, joint self-attention weights are obtained to identify the crucial features of houses for buyers.

A. DATA DESCRIPTION AND PREPROCESSING

To train the model and verify its effectiveness, we use actual selling prices listed in real estate transaction data; public facility data such as hospitals, parks, schools, convenience stores, and MRT stations; and satellite maps.

1) DATA DESCRIPTION

The house transaction dataset is collected from 2017 to 2018 for transactions in Taipei City and New Taipei City.

The dataset includes a total of 93,696 transactions classified into three types: only building, land with building, and land with building and parking space. We filter the data to exclude undesired data types and outliers. After the filtration process, 92,857 transaction records remain, and we focus on unit prices, which constitute the target of our study.

For public facility data, We acquired the open access data from platforms maintained by the Taipei City and New Taipei City governments. These datasets include lists of parks, hospitals, schools, convenience stores, and MRT stations. The column on parks contains data concerning the name, latitude–longitude coordinates, and address of each park as well as information on transportation (i.e., how to reach the park) and sports facilities. The column on hospitals contains information on the name and address of each hospital as well as the contact information of each. The column on schools contains information on institution type (elementary school, junior or senior high school, university), postal code, address, and latitude–longitude coordinates. The column on convenience stores and MRT stations provides latitude–longitude coordinates. All public facilities, including parks, hospitals, schools, convenience stores, and MRT stations, are registered on the platforms maintained by the Taipei City and New Taipei City governments. Therefore, missing data was not a problem in the present study. The only limitation concerning the accuracy of the data was that the addresses of the facilities were not exact. For example, an address may appear as “No. 296 to Lane 304, Section 3, Xinglong Road (in front of the Marine Patrol Department).” To resolve this problem, we used a Google API to crawl the coordinates of each address. Table 1 presents the statistics of the public facilities. Convenience stores are particularly numerous.

TABLE 1. Statistics of public facility data.

Public facilities	Numbers
Convenience store	3719
Hospital	128
Park	990
School	665
MRT station	131

For satellite map data, we utilize the Google Maps API to crawl satellite maps (640×640) according to given house addresses. The satellite maps present the terrain surrounding houses that can indicate whether a house is near a green space, mountains, numerous buildings, or a coastline. Because they have abundant information, these maps are treated as part of the input for our model to supplement the obtained house information. Fig. 7 depicts examples of the assessed satellite images. Each house has different surrounding environments. In Fig. 7a, mountains are on the right-hand side of the houses; we can speculate this building is located in the mountains. Fig. 7b presents houses near the sea that may be located along the coastline of Northern Taiwan.

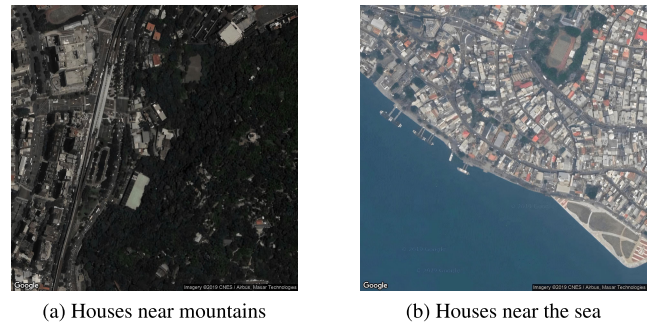


FIGURE 7. Examples satellite images.

2) DATA PREPROCESSING

After obtaining our house transaction dataset, we preprocess the data according to different features. The dataset comprises a total of 28 columns, and we adopt only some of these columns as our input features in the prediction model. We preprocess only these special columns, the corresponding preprocessing methods are presented herein. We adopt appropriate normalization methods to manage the other columns. This section introduces the method of processing the following columns: land sector position, building sector, and house address plaque. The columns represent house locations, but the transaction data do not have records of exact addresses. For example, an address may appear as “No. 1–No. 30 Kangding Rd., Wanhua Dist., Taipei City, 10843, Taiwan (R.O.C.).” Because of this limitation, we employ the Google Static API to crawl the latitudes and longitudes of the address by using two methods:

- Center section: In the aforementioned example address, the center GPS coordinate between No. 1 and No. 30 is selected as the house address
- Division of section into four different addresses: We split the address into Nos. 1, 11, 21, and 30 and crawl the corresponding addresses as possible house locations.

To map our attention model, we classify the aforementioned data into 13 attribute groups according to their properties. The resulting data are among the inputs in the attention models. The 13 attribute groups are as follows.

- Total area: This group comprises the areas (in square meters) of land, buildings, and berths. Because these three features are related to area, we classify them into this group.
- Pattern: This group involves a building’s patterns, namely rooms, halls, health, and compartments. These attributes represent the pattern of a house and are thus classified into this group.
- Age: This group comprises house age and transaction time score. Because these two features are transformed using the transaction date column, we classify them into this group.
- Number: This group involves attributes recorded in the following column: transaction pen number. Generally, if a house occupies several plot numbers, it may have

a large land area. This type of house may belong to a condominium with a large community.

- **Type:** This group comprises the building state.
- **Location:** This group involves the coordinates of one of the four separated addresses and its corresponding district.
- **Floor:** This group comprises the floor level of a house and the total number of floors in a building; it indicates whether a house is located on a high or low floor.
- **Store, hospital, park, school, and MRT:** These groups comprise the numbers of these facilities near a house and can represent the level of life functions.
- **Map:** This group contains the satellite map features after extraction with STN.

B. EVALUATION METRICS

This section provides an overview of evaluation metrics used to measure model performance. To evaluate the model performance, the MAPE is calculated as the error rate and can be expressed as follows:

$$MAPE = \frac{1}{n} \sum \frac{|Price_{pred} - Price_{true}|}{Price_{true}} \quad (7)$$

$Price_{pred}$ represents the price predicted by the model and $Price_{true}$ represents the actual selling price recorded in the house transaction data. We derive the error for each sample and then averaged the total errors for n samples to obtain the final error rate, which is used to evaluate our model performance. We separate the data of Taipei City and New Taipei City into 15 regions on the basis of the divisions made by the land administration office. We then use the data to obtain the experimental results for these 15 regions and the overall error.

C. HOUSE TRANSACTION AND HETEROGENEOUS DATA

In the first experiment, we hypothesize that house transaction data alone cannot be used to precisely predict unit prices and that supplemental data (in addition to house information) are required to improve prediction accuracy.

To verify the aforementioned hypothesis, we adopt various models by importing different datasets; we compare results that are obtained using machine learning and deep learning models. Several approaches are employed to verify the hypothesis, and the following items constitute a baseline (control) model in the experiment.

- **Baseline:** The unit price of the house that is nearest to house A is used as the predicted price of house A. This operation is repeated for all samples, and the MAPE is calculated. This baseline operation is executed to verify whether the machine learning and deep learning models can learn rules from transaction data.
- **LR 1, XGB 1, LGB 1:** Only one address and other features are employed to predict price using linear regression, XGBoost, and Light Gradient Boosted Machine (LightGBM) methods. These methods are executed by importing the sklearn package.

- **LR 4, XGB 4, LGB 4:** Four separate addresses are used to generate four prices, which are then averaged to obtain the predicted price through the linear regression, XGBoost, and LightGBM methods.
- **LR f, XGB f, LGB f:** The four separate addresses, house transaction data, and corresponding public facility data are applied to predict a price using the three previously mentioned models.
- **FC 1, FC 4:** Only one address and four addresses with other house attributes are employed for price prediction using models with several fully connected layers.
- **FC f:** Public facility data corresponding to four addresses are imported for price prediction using models with several fully connected layers.
- **FC sf:** Heterogeneous data comprising satellite maps and public facility data that correspond to four addresses are imported for price prediction using a CNN model with fully connected layers. The MAPE is calculated to determine model performance.
- **STN sf:** The **FC f** model is improved by introducing an STN [7] to extract the most crucial features from the satellite maps for price prediction.

According to the preceding controls, most models are trained with a gated neural network, which is introduced in Section III-E. As mentioned, using only the center point of a section of road as the house address cannot fully represent the house location. Therefore, the road section can be separated into four addresses to increase the number of possible house addresses. After the four separate house addresses are generated, the corresponding data are input into the gated neural network to identify possible house addresses. This is the main method for most of the prediction models.

Table 2 presents a comparison of experimental results obtained using the machine learning models with the baseline model. The y -axis represents all approaches used in the experiment, and the x -axis represents all targets used for evaluation. Bold font indicates the lowest error rate among the methods for each column. Most of the adopted machine learning models, except for the linear regression model, have a higher performance in terms of the MAPE than does the baseline model. Linear regression is performed using statistical information to identify the relationship between dependent variables and multiple independent variables. In the linear regression model used in this paper, a simple linear regression approach is adopted to fit overall house prices. However, because the real estate market is complicated, this model alone cannot be used to fit all the prices. Consequently, the linear regression model produces the least favorable results. By contrast, XGBoost has the lowest MAPE, regardless of whether only one address, four separate addresses, or public facility data are used. The results in Table 2 prove that importing additional data to supplement house information can improve prediction accuracy.

TABLE 2. Performance comparison with different input by adopting traditional machine learning methods.

Measurement	Xizhi	Sanchong	Shulin	Xindian	Tamsui	Banjiao	Xinzhuang	Ruifang	Zhonghe	Shilin	Zhongshan	Jiancheng	Guting	Songshan	Daan	Avg.
rule_based	16.15%	24.71%	19.85%	21.36%	12.53%	22.05%	18.36%	54.26%	24.44%	28.43%	27.79%	34.22%	27.44%	35.11%	29.45%	23.82%
LR_1	123.86%	27.44%	43.98%	55.10%	47.03%	51.02%	29.84%	781.71%	28.64%	42.92%	29.85%	30.97%	32.60%	33.16%	43.75%	43.34%
XGB_1	10.97%	14.23%	15.09%	14.27%	12.13%	12.92%	11.13%	65.59%	16.88%	18.25%	16.91%	21.38%	16.90%	22.53%	17.83%	15.45%
LGB_1	13.78%	14.99%	18.14%	17.03%	14.29%	14.79%	13.10%	52.84%	17.88%	19.12%	17.92%	21.59%	17.91%	23.73%	19.41%	17.00%
LR_4	123.80%	27.20%	44.36%	54.83%	46.62%	51.28%	29.77%	778.87%	28.63%	43.00%	29.85%	30.96%	32.60%	33.19%	43.77%	43.32%
XGB_4	10.24%	13.84%	14.76%	14.21%	10.74%	12.29%	10.90%	45.49%	16.30%	17.19%	16.76%	20.65%	16.02%	22.13%	18.36%	14.87%
LGB_4	13.47%	14.89%	17.89%	16.51%	14.32%	14.83%	12.98%	51.04%	17.79%	19.11%	17.80%	21.56%	17.74%	23.53%	19.08%	16.87%
LR_f	33.42%	21.88%	29.82%	29.56%	28.24%	46.64%	19.95%	72.43%	21.30%	38.73%	29.52%	35.08%	30.05%	30.30%	29.65%	30.41%
XGB_f	10.23%	13.74%	14.25%	13.55%	10.16%	12.05%	10.84%	45.33%	16.18%	16.49%	15.85%	19.83%	15.48%	21.80%	17.80%	14.46%
LGB_f	13.10%	13.81%	17.46%	15.52%	13.03%	14.65%	12.90%	54.26%	17.36%	18.81%	17.23%	21.01%	17.55%	22.98%	19.00%	16.40%

TABLE 3. Performance comparison for deep learning models.

Measurement	Xizhi	Sanchong	Shulin	Xindian	Tamsui	Banjiao	Xinzhuang	Ruifang	Zhonghe	Shilin	Zhongshan	Jiancheng	Guting	Songshan	Daan	Avg.
rule_based	16.15%	24.71%	19.85%	21.36%	12.53%	22.05%	18.36%	54.26%	24.44%	28.43%	27.79%	34.22%	27.44%	35.11%	29.45%	23.82%
LR_1	123.86%	27.44%	43.98%	55.10%	47.03%	51.02%	29.84%	781.71%	28.64%	42.92%	29.85%	30.97%	32.60%	33.16%	43.75%	43.34%
XGB_1	10.97%	14.23%	15.09%	14.27%	12.13%	12.92%	11.13%	65.59%	16.88%	18.25%	16.91%	21.38%	16.90%	22.53%	17.83%	15.45%
LGB_1	13.78%	14.99%	18.14%	17.03%	14.29%	14.79%	13.10%	52.84%	17.88%	19.12%	17.92%	21.59%	17.91%	23.73%	19.41%	17.00%
LR_4	123.80%	27.20%	44.36%	54.83%	46.62%	51.28%	29.77%	778.87%	28.63%	43.00%	29.85%	30.96%	32.60%	33.19%	43.77%	43.32%
XGB_4	10.24%	13.84%	14.76%	14.21%	10.74%	12.29%	10.90%	45.49%	16.30%	17.19%	16.76%	20.65%	16.02%	22.13%	18.36%	14.87%
LGB_4	13.47%	14.89%	17.89%	16.51%	14.32%	14.83%	12.98%	51.04%	17.79%	19.11%	17.80%	21.56%	17.74%	23.53%	19.08%	16.87%
LR_f	33.42%	21.88%	29.82%	29.56%	28.24%	46.64%	19.95%	72.43%	21.30%	38.73%	29.52%	35.08%	30.05%	30.30%	29.65%	30.41%
XGB_f	10.23%	13.74%	14.25%	13.55%	10.16%	12.05%	10.84%	45.33%	16.18%	16.49%	15.85%	19.83%	15.48%	21.80%	17.80%	14.46%
LGB_f	13.10%	13.81%	17.46%	15.52%	13.03%	14.65%	12.90%	54.26%	17.36%	18.81%	17.23%	21.01%	17.55%	22.98%	19.00%	16.40%
FC_1	11.25%	13.93%	14.41%	17.54%	9.82%	12.23%	11.24%	32.13%	16.18%	19.02%	17.89%	21.20%	18.23%	23.55%	19.39%	15.52%
FC_4	10.75%	13.82%	14.49%	16.72%	9.25%	12.14%	10.94%	33.87%	15.75%	18.36%	17.11%	20.76%	17.95%	23.01%	18.98%	15.12%
FC_f	9.61%	13.24%	12.90%	13.54%	8.55%	11.38%	10.13%	27.86%	15.31%	16.75%	17.13%	19.20%	16.33%	21.69%	20.76%	14.17%
FC_sf	9.80%	13.27%	12.54%	12.83%	8.27%	10.93%	9.71%	28.09%	15.16%	16.09%	15.90%	19.30%	15.98%	21.85%	20.16%	13.79%
STN_sf	9.20%	12.82%	12.38%	12.87%	7.90%	10.86%	9.36%	28.75%	15.46%	16.42%	15.41%	19.88%	16.32%	20.91%	20.36%	13.63%

We adopt deep learning models under the same experimental settings to compare their performance with the machine learning models (Table 3). We note that introducing data regarding satellite maps and public facilities such as parks, schools, and hospitals can reduce the overall error rate for **FC 4**, and **FC s**. We also apply the STN to the satellite maps to improve **FC s**, and **STN s** outperforms the other approaches. **STN s** yields the smallest MAPE for 8 of the 15 regions and outperforms the other models in terms of the overall region error rate. The MAPEs derived for some regions, such as Ruifang, are high. Analyzing such regions reveals that the transaction data in these areas rarely fit the model. Additionally, these regions have lower house prices, which may increase the error rate relative to that in areas with higher house prices. All models adopting the gated neural network for prediction outperform the models adopting only the center point as the house address. These experimental results prove that importing heterogeneous data can improve model performance and that deep learning models can outperform machine learning models such as XGBoost and LightGBM.

D. JOINT SELF-ATTENTION COMPARED WITH OTHER MODELS

We conduct another experiment to verify whether adopting an attention mechanism can improve experimental results. Different mechanisms should be used to analyze different house types. For example, two houses may have similar patterns, building types, and shifting total areas but different ages or locations. Therefore, different attention weights should be used to predict their prices according to their situations, and their prices should not be fit using the same model.

Accordingly, we use different attention mechanism-based models to validate our hypothesis. Except for the previous experimental models and settings, all attention models used in this experiment are listed as follows:

- **Gated, Attention, Self-Attention, and Joint Self-Attention:** We adopt these models as attention mechanisms to derive model weights. The model inputs are the same as those used in the previous experiment (i.e., the previously mentioned experimental settings).

TABLE 4. Performance comparison with attention mechanism.

Measurement	Xindian	Tamsui	Banqiao	Zhonghe	Shilin	Zhongshan	Jiancheng	Guting	Songshan	Daan	Avg.
rule_based	21.36%	12.53%	22.05%	24.44%	28.43%	27.79%	34.22%	27.44%	35.11%	29.45%	25.34%
LR_1	55.10%	47.03%	51.02%	28.64%	42.92%	29.85%	30.97%	32.60%	33.16%	43.75%	38.86%
XGB_1	14.27%	12.13%	12.92%	16.88%	18.25%	16.91%	21.38%	16.90%	22.53%	17.83%	15.60%
LGB_1	17.03%	14.29%	14.79%	17.88%	19.12%	17.92%	21.59%	17.91%	23.73%	19.41%	16.78%
LR_4	54.83%	46.62%	51.28%	28.63%	43.00%	29.85%	30.96%	32.60%	33.19%	43.77%	35.29%
XGB_4	14.21%	10.74%	12.29%	16.30%	17.19%	16.76%	20.65%	16.02%	22.13%	18.36%	15.00%
LGB_4	16.51%	14.32%	14.83%	17.79%	19.11%	17.80%	21.56%	17.74%	23.53%	19.08%	16.71%
LR_f	29.56%	28.24%	46.64%	21.30%	38.73%	29.52%	35.08%	30.05%	30.30%	29.65%	29.98%
XGB_f	13.55%	10.16%	12.05%	16.18%	16.49%	15.85%	19.83%	15.48%	21.80%	17.80%	14.55%
LGB_f	15.52%	13.03%	14.65%	17.36%	18.81%	17.23%	21.01%	17.55%	22.98%	19.00%	16.26%
FC_1	17.54%	9.82%	12.23%	16.18%	19.02%	17.89%	21.20%	18.23%	23.55%	19.39%	15.56%
FC_4	16.72%	9.25%	12.14%	15.75%	18.36%	17.11%	20.76%	17.95%	23.01%	18.98%	15.13%
FC_f	13.54%	8.55%	11.38%	15.31%	16.75%	17.13%	19.20%	16.33%	21.69%	20.76%	14.44%
FC_sf	12.83%	8.27%	10.93%	15.16%	16.09%	15.90%	19.30%	15.98%	21.85%	20.16%	14.06%
STN_sf	12.87%	7.90%	10.86%	15.46%	16.42%	15.41%	19.88%	16.32%	20.91%	20.36%	14.02%
Gated	13.01%	8.09%	10.97%	15.15%	16.29%	15.86%	19.38%	15.86%	21.33%	18.41%	13.93%
Attention	12.89%	8.24%	11.13%	14.78%	16.19%	15.52%	19.08%	15.66%	20.77%	19.44%	13.82%
Self Attention	13.44%	8.28%	11.24%	15.13%	16.55%	15.26%	18.54%	15.91%	21.12%	18.31%	13.87%
Joint Self-Attention	12.92%	8.42%	10.70%	14.82%	15.98%	15.41%	18.37%	15.23%	20.83%	18.00%	13.61%
Gated_sf	13.15%	8.44%	11.55%	15.01%	15.26%	15.29%	18.77%	16.15%	21.78%	19.48%	13.94%
Attention_sf	13.02%	8.53%	11.20%	15.04%	16.12%	15.70%	18.35%	15.19%	20.92%	18.29%	13.80%
Self-Attention_sf	13.51%	8.17%	10.77%	14.91%	15.79%	15.68%	18.47%	15.64%	21.24%	17.82%	13.69%
Joint Self-Attention_sf	12.72%	8.29%	10.74%	14.80%	15.67%	15.24%	18.04%	15.31%	20.51%	17.74%	13.49%

The given transaction sample is used as the query to employ the attention mechanism.

- **Gated sf, Attention sf, Self Attention sf, and Joint Self-Attention sf:** These models have the same framework as the four aforementioned models. Nevertheless, they have an additional input, namely map attributes.

Table 4 presents the experimental results; the meanings of the y -axis and x -axis are same as those in Tables 2 and 3. Because the number of data entries in each dataset is different, and because the resolutions of some of the map attributes were quite low, thereby affecting the training performance, we divided them into six categories, indicating the best result in each category in bold. The performance of joint self-attention in the last two categories was noted in more than half of the area. Moreover, the average performance was better than that of all the other benchmarks. When the input data comprise only the house attributes and public facilities, the models with the attention mechanisms outperform **FC f**. Specifically, our proposed joint self-attention model, **Joint Self-Attention**, has an MAPE of 13.61%. When all features, including house attributes, public facilities, and satellite map data, are input, most of the attention-based models

outperform **STN sf**. Our proposed **Joint Self-Attention sf** model, which involves a joint self-attention architecture and an STN image feature extractor, has the lowest MAPE (13.49) among the compared models. Moreover, several regions are associated with low MAPEs because they have more house transaction data and similar house price distributions. This experiment demonstrates that the adoption of an attention mechanism that can assign different weights according to different samples can increase model flexibility in predicting house prices. The bottom part of Table 4 indicates that model performance improves after the adoption of the attention mechanism. Moreover, when the joint self-attention mechanism is imported, the model can learn the two-hop interaction between two attributes to obtain precise attention weights; the joint self-attention mechanism is associated with superior performance compared with the other attention mechanisms.

E. JOINT SELF-ATTENTION COMPARED WITH OTHER ATTENTION MECHANISM

To verify the effectiveness of the joint self-attention mechanism, we conducted an ablation experiment. The data used in this experiment were the original inputs with map attributes.

TABLE 5. Performance comparison with state-of-the-art attention mechanism.

Measurement	Xizhi	Sanchong	Shulin	Xindian	Tamsui	Banqiao	Xinzhuang	Ruifang	Zhonghe	Shilin	Zhongshan	Jiancheng	Guting	Songshan	Daan	Avg.
Self-Attention_sf	9.26%	12.90%	12.30%	13.51%	8.17%	10.77%	9.91%	27.15%	14.91%	15.79%	15.68%	18.47%	15.64%	21.24%	17.82%	13.52%
Co-attention_sf	9.54%	12.89%	13.24%	12.33%	8.66%	13.01%	9.82%	29.57%	14.91%	15.73%	16.01%	18.76%	16.51%	21.18%	18.26%	13.77%
CBAM_sf	9.17%	13.12%	12.11%	13.25%	8.45%	11.30%	9.87%	24.05%	14.80%	16.16%	15.89%	18.65%	15.03%	21.10%	17.59%	13.57%
Multihead Attention_sf	Self-10.07%	13.39%	12.77%	13.31%	8.65%	11.58%	10.28%	32.50%	15.17%	16.22%	15.51%	20.00%	16.26%	22.20%	19.01%	14.01%
Joint Self-Attention_sf	9.12%	12.67%	12.44%	12.72%	8.29%	10.74%	9.82%	29.32%	14.80%	15.67%	15.24%	18.04%	15.31%	20.51%	17.74%	13.31%

We applied heterogeneous attention-based approaches in our control groups. The attention models used in this experiment are listed as follows:

- **Co-attention sf:** Co-attention networks are adept at finding information that is common among various sets of multimodal data. The co-attention mechanism builds attention blocks according to each modality and incorporates the heterogeneous information into the final decisions.
- **CBAM sf:** This convolutional block attention module (CBAM) achieved excellent results on the image classification task. We applied it to capture key image features.
- **Multihead Self-Attention sf:** Unlike the original self-attention mechanism, this model separates each query, key, and value into multiple sets and applies the attention mechanism to each. The advantage of this approach is that it makes each head concentrate on the task and learn information in different locations both efficiently and effectively.

Table 5 presents the experiment results. Our proposed joint self-attention model achieved the lowest average MAPE of 13.31%. The multihead self-attention_sf had the highest MAPE, most likely because the data characteristics were not suitable to this method. As mentioned, the multihead self-attention model divides each query, key, and value into several pieces. It is appropriate for use in NLP tasks but may not be applicable to the present situation. Notably, the co-attention sf, CBAM_sf, and multihead self-attention_sf did not perform as well as the attention_sf or the self-Attention_sf, perhaps because the co-attention_sf, CBAM_sf, and multihead self-attention_sf are typically applied in visual question answering tasks, classification tasks, and NLP tasks, respectively. Moreover, the present study was focused on a price prediction problem. In essence, although all these models use attention mechanisms, the method may not achieve satisfactory performance on other tasks.

F. APPLICATION TO HOUSE PRICE ANALYSIS IN METROPOLITAN REGIONS

The third experiment focuses on the metropolitan regions of Taichung City and Kaohsiung City. Specifically, we analyze the real estate markets and predict house prices in these regions to observe whether the real estate markets in these major cities have similar trends to those of Taipei and New Taipei City. We collect for the same period as those for the first experiment and then filter them. A total

of 58,425 transactions (47,000 for training and 11,425 for testing) are collected for Taichung, and 49,551 transaction records (40,000 for training and 9,551 for testing) are collected for Kaohsiung.

In contrast to the previous experimental settings, the present experiment involves the use of only public facility data and transaction data for model training and price prediction. Public facility data are acquired from the Taichung and Kaohsiung governments. In addition, we use train stations to represent public transportation in Taichung because the city's MRT system has not yet opened. All the features of the house transaction data are preprocessed using the methods described in Section IV-C. After preprocessing, we use the settings of the first two experiments to execute house price analysis for the two metropolitan regions.

Tables 6 and 7 present the experimental results for Taichung and Kaohsiung, respectively. The results confirm our hypothesis that importing more data can improve prediction accuracy; the results demonstrate that the deep learning models yield lower MAPEs than do the machine learning models. The MAPE decreases considerably when public facility data are imported, indicating that public facilities greatly influence house prices. Moreover, we use attention mechanisms in this experiment (lower parts of the tables) to verify that such mechanisms can automatically assign different weights to each feature based on data type and can identify crucial features for different buyers. Our proposed **Joint Self-Attention** model outperforms all other models, including the machine learning, deep learning, and other attention models. Specifically, **Joint Self-Attention** yields the lowest MAPE for 5 of the 11 regions of Taichung City. For Kaohsiung City, **Joint Self-Attention** outperforms the other models in that it yields the lowest MAPE for most of the 12 regions of the city.

The experimental results for these two metropolitan cities reveal that the error rates for these cities are similar to those for Taipei and New Taipei City. Specifically, the error rates are 12.11% for Taichung City, 15.89% for Kaohsiung City, and 13.31% for Taipei City and New Taipei City. The error rate for Kaohsiung City is higher than those for the other cities because the house price distribution in Kaohsiung City is relatively concentrated, whereas the other cities have similar house price distributions. The MAPEs derived for individual regions indicate that some regions may have insufficient data and low house prices, which may lead to an increased MAPE relative to that for other regions with relatively high house prices.

TABLE 6. Apply joint self-attention on housing price issue in Taichung.

Measurement	Zhongshan	Zhongzheng	Zhongxing	Fengyuan	Dajia	Qingshui	Dongshih	Yatan	Dali	Taiping	Longjing	Avg.
rule_based	24.86%	23.18%	19.89%	25.81%	48.27%	20.54%	40.81%	25.86%	26.66%	17.92%	26.44%	23.11%
LR_1	29.24%	24.74%	23.02%	25.38%	42.87%	38.75%	60.00%	20.52%	39.47%	21.37%	55.59%	27.81%
XGB_1	15.17%	12.86%	11.69%	18.01%	31.73%	13.61%	21.79%	14.34%	16.18%	11.66%	23.86%	14.03%
LGB_1	16.01%	14.10%	12.66%	19.08%	32.89%	15.06%	29.04%	15.29%	17.39%	13.10%	23.07%	15.17%
LR_4	29.22%	24.72%	23.01%	25.33%	43.02%	38.70%	60.21%	20.54%	39.50%	21.40%	55.70%	27.80%
XGB_4	14.93%	12.54%	11.21%	17.39%	28.42%	12.70%	21.72%	13.91%	15.81%	11.09%	21.24%	13.52%
LGB_4	15.86%	13.95%	12.60%	18.76%	32.19%	14.50%	27.98%	15.10%	17.40%	13.05%	22.73%	15.01%
LR_f	24.42%	22.99%	22.21%	24.31%	40.65%	33.95%	72.35%	19.41%	34.73%	19.94%	34.83%	25.24%
XGB_f	14.83%	12.33%	10.81%	17.14%	27.82%	12.52%	21.73%	13.23%	15.59%	10.82%	20.28%	13.23%
LGB_f	15.69%	13.74%	12.15%	18.85%	33.04%	14.61%	26.47%	14.19%	17.18%	12.50%	23.24%	14.74%
FC_1	15.19%	12.92%	12.89%	17.14%	22.69%	12.10%	25.74%	14.60%	16.56%	11.90%	19.29%	14.06%
FC_4	15.20%	12.63%	12.22%	17.59%	22.32%	12.16%	21.06%	14.20%	16.35%	11.54%	21.21%	13.79%
FC_f	14.63%	12.17%	10.90%	16.65%	25.68%	11.24%	23.60%	12.75%	16.14%	10.35%	19.96%	13.03%
Gated	14.56%	11.65%	10.54%	15.38%	22.09%	10.91%	21.31%	12.34%	14.65%	9.02%	17.57%	12.33%
Attention	13.90%	11.59%	10.39%	15.54%	23.12%	10.84%	20.82%	12.38%	14.59%	9.19%	16.38%	12.20%
Self Attention	13.48%	11.53%	10.40%	15.43%	23.34%	11.07%	21.42%	12.41%	15.20%	9.80%	16.87%	12.26%
Joint Self-Attention	13.54%	11.50%	10.23%	15.28%	22.07%	10.86%	23.09%	12.31%	14.25%	9.71%	17.19%	12.11%

G. SYSTEM DEMONSTRATION

In this experiment, we visualize our joint self-attention model and identify features with high weights. Additionally, to determine public facilities around a house, we use a representation method to display the number of hospitals or parks near a house. We consider one case for demonstration. In this case, unit prices are in the range of US\$200,000 to US\$210,000. Additionally, transaction data, including basic house information, the five most crucial features calculated by the joint self-attention model, the surrounding environment, and the recommendation reasons, are provided in the following report; the data are described in detail.

House Recommendation Report

House Information Introduction

- District: Wanhua
- Address: No. 1–30, Ln. 198, Wanda Rd., Wanhua Dist., Taipei City
- Construction completion date: November 23, 2016
- Transaction date: August 16, 2017
- Building state: suite
- Transaction pen number: 1 Land, 1 Building, 0 Berth
- Land shifting area: 10.85
- Building area: 35.05
- Berth area:
- Building pattern, room: 1
- Building pattern, hall: 1
- Building pattern, health: 1
- Building pattern, compartment: No
- Floor level: 5

- Total floor: 5
- Management organization: No
- Unit price: 205,421

House Price Predicted by the Proposed Model

- Predicted price: US\$185,418
- Unit price: US#205,421
- MAPE: 4.9%

Five Most Crucial Attributes Recommended by the Proposed Model

- Convenience store attributes
- Park attributes
- Age attributes with park attributes
- Age attributes with school attributes
- Floor attributes with convenience store attributes

Surrounding Public Facilities

- 20 convenience stores, 1 park, and 1 school within 1 km in Quadrant I
- 11 convenience stores, 8 parks, and 7 schools within 1 km in Quadrant II
- 24 convenience stores, 5 parks, and 4 schools within 1 km in Quadrant III
- 15 convenience stores, 3 parks, and 1 school within 1 km in Quadrant IV

Recommended Target Buyers

- Unmarried office workers with exercise habits may intend to buy this type of house

TABLE 7. Apply joint self-attention on housing price issue in Kaohsiung.

Measurement	Yancheng	Sinsing	Cainjhen	Nanzih	Sanmin	Gangshan	Fongshan	Cishan	Renwu	Lujhu	Meinong	Daliao	Avg.
rule_based	29.32%	31.21%	41.48%	27.09%	24.86%	59.33%	24.33%	92.28%	26.62%	88.60%	115.50%	89.71%	34.51%
LR_1	18.14%	24.57%	21.44%	16.97%	17.31%	31.69%	17.08%	41.01%	16.90%	25.99%	109.34%	38.45%	20.58%
XGB_1	19.92%	24.66%	23.66%	18.77%	18.84%	32.10%	18.83%	45.51%	19.35%	34.28%	93.10%	32.07%	21.96%
LGB_1	29.32%	31.19%	41.48%	27.08%	24.86%	59.47%	24.39%	92.67%	26.62%	88.47%	116.56%	89.67%	34.52%
LR_4	17.44%	23.63%	20.10%	16.75%	17.08%	31.09%	16.81%	44.49%	16.75%	25.41%	92.04%	33.96%	19.88%
XGB_4	19.78%	24.74%	22.99%	18.91%	18.68%	32.22%	18.46%	44.97%	19.06%	34.20%	94.75%	31.58%	21.79%
LGB_4	25.70%	32.16%	29.90%	23.73%	28.34%	46.26%	24.68%	68.17%	23.74%	52.97%	40.24%	48.42%	29.10%
LR_f	16.82%	23.56%	18.23%	15.64%	16.86%	28.77%	15.92%	38.41%	15.66%	26.46%	99.69%	30.52%	18.86%
XGB_f	19.22%	24.30%	22.01%	18.09%	18.16%	30.63%	18.43%	41.62%	18.38%	31.54%	87.03%	32.20%	21.14%
LGB_f	34.02%	47.41%	34.60%	27.85%	33.74%	48.57%	29.00%	57.27%	28.16%	96.85%	49.29%	53.86%	35.64%
FC_1	18.25%	21.85%	15.66%	16.89%	17.91%	30.73%	15.07%	36.01%	17.82%	25.81%	51.79%	26.07%	18.73%
FC_4	17.96%	20.77%	15.46%	17.02%	18.19%	26.88%	15.15%	33.57%	16.46%	21.90%	47.68%	24.13%	18.16%
FC_f	16.28%	18.97%	15.09%	14.96%	16.45%	25.64%	14.14%	29.89%	14.30%	24.98%	37.49%	24.04%	16.73%
Gated	15.62%	18.17%	14.86%	14.14%	15.99%	24.35%	14.31%	28.19%	14.50%	24.33%	36.54%	23.10%	16.25%
Attention	15.29%	18.51%	14.59%	14.04%	16.44%	23.22%	14.31%	33.12%	13.36%	24.39%	38.22%	21.96%	16.09%
Self Attention	15.41%	19.14%	14.27%	13.77%	16.47%	23.48%	13.99%	28.34%	13.40%	22.15%	39.66%	21.98%	15.98%
Joint Self-Attention	15.13%	18.48%	14.36%	14.24%	15.87%	23.94%	13.99%	28.39%	12.79%	22.90%	37.59%	21.94%	15.89%

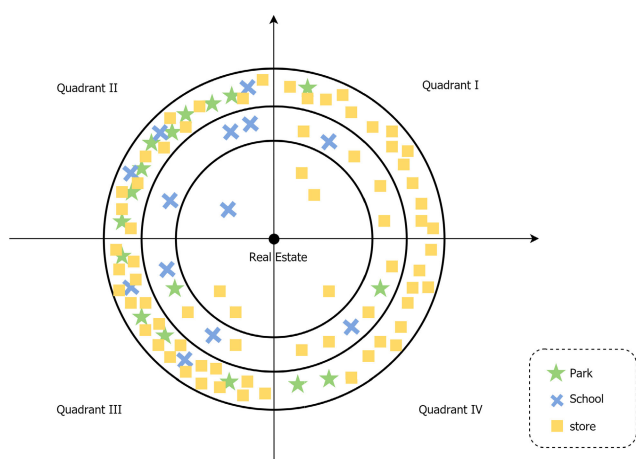


FIGURE 8. Public facilities surrounding the house.

In this case, the five most crucial attributes include several public facility attributes and house features such as house age and floor level. We then focus on these five attributes predicted by our proposed model and analyze the transaction data and surrounding environment. The transaction data reveal that the house is approximately 1 year old and located on the top floor. The building type is a suite. This information indicates that the buyer sought a new house on a high floor. The surrounding environment plotted in Fig. 8 indicates several schools, parks, and convenience stores near the house, particularly in the upper left-hand corner and lower right-hand corner. These features indicate that the buyer who purchased this house not only considered the house age and floor level but also surrounding amenities. We speculate the

buyer to be a single office worker who buys food at night from convenience stores. The numerous parks nearby enable the buyer to take walks after work or on weekends. Although the age and floor attributes are not among the five major features, the public facilities of stores and parks can increase purchase intention. In summary, we conclude that unmarried office workers may intend to buy this type of house.

V. CONCLUSION

In this study, we develop a robust model for effective house prediction. In this model, we import heterogeneous data to supplement house information and propose an attention mechanism to automatically assign weights according to different features or samples. We constructed an end-to-end model based on STN techniques and extended the concept of self-attention mechanism to propose a joint-self attention mechanism. In addition, we compare several attention mechanisms in our experiment to verify the effectiveness of the proposed method. First, to our hypothesis that external information can improve prediction results, we compare results obtained from machine learning and deep learning models. Second, we adopt different attention models to prove that attention mechanisms can improve prediction accuracy. Our proposed joint self-attention model, which considers two-hop relationships between different attributes, exhibits particularly high performance levels. Third, to verify the effectiveness of our proposed model, we applied it to analyze the real estate markets in Taichung City and Kaohsiung City. Finally, we demonstrate the performance of our proposed model by using a case to explain why and what types of buyers would select and purchase a specific house. The main contributions of this work include the development of a model

that imports heterogeneous data to supplement house information, the adoption of an STN to process satellite maps in order to extract image features, and the development of a joint self-attention model for learning two-hop information between attributes. On the basis of the advantages of these methods, our proposed model outperforms other models in all experiments.

Future work can import additional information, such as construction companies, building contractors, and undesirable facilities, that may affect buyer intentions and house prices. In addition, the data limitations of this study should be improved to represent specific features more precisely. For example, the transaction data do not contain precise house addresses, and we cannot obtain complete public facility data. Moreover, we cannot retrieve images of house interiors to consider the interior design of the houses. Finally, satellite maps can be adopted along with saliency maps to visualize influential elements for price prediction. If these aspects can be improved, prediction results will be more precise.

REFERENCES

- [1] R. Gupta, A. Kabundi, and S. M. Miller, "Forecasting the US real house price index: Structural and non-structural models with and without fundamentals," *Econ. Model.*, vol. 28, no. 4, pp. 2013–2021, Jul. 2011.
- [2] J. Mu, F. Wu, and A. Zhang, "Housing value forecasting based on machine learning methods," *Abstract Appl. Anal.*, vol. 2014, pp. 1–7, Aug. 2014.
- [3] L. Bork and S. V. Møller, "Forecasting house prices in the 50 states using dynamic model averaging and dynamic model selection," *Int. J. Forecasting*, vol. 31, no. 1, pp. 63–78, Jan. 2015.
- [4] A. Ng and M. Deisenroth, "Machine learning for a London housing price prediction mobile application," Imperial College London, London, U.K., 2015.
- [5] M. Risse and M. Kern, "Forecasting house-price growth in the euro area with dynamic model averaging," *North Amer. J. Econ. Finance*, vol. 38, pp. 70–85, Nov. 2016.
- [6] B. Afonso, L. Melo, W. Oliveira, S. Sousa, and L. Berton, "Housing prices prediction with a deep learning and random forest ensemble," in *Proc. Anais do 16th Encontro Nacional de Inteligência Artif. e Computacional*, 2019, pp. 389–400.
- [7] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.
- [8] Z. Peng, Q. Huang, and Y. Han, "Model research on forecast of second-hand house price in Chengdu based on XGboost algorithm," in *Proc. IEEE 11th Int. Conf. Adv. Infocomm Technol. (ICAIT)*, Oct. 2019, pp. 168–172.
- [9] C. R. Madhuri, G. Anuradha, and M. V. Pujitha, "House price prediction using regression techniques: A comparative study," in *Proc. Int. Conf. Smart Struct. Syst. (ICSSS)*, Mar. 2019, pp. 1–5.
- [10] T. D. Phan, "Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia," in *Proc. Int. Conf. Mach. Learn. Data Eng. (iCMLDE)*, Dec. 2018, pp. 35–42.
- [11] A. S. Temür, M. Akgün, and G. Temür, "Predicting housing sales in Turkey using ARIMA, LSTM and hybrid models," *J. Bus. Econ. Manage.*, vol. 20, no. 5, pp. 920–938, Jul. 2019.
- [12] L. Yu, C. Jiao, H. Xin, Y. Wang, and K. Wang, "Prediction on housing price based on deep learning," *Int. J. Comput. Inf. Eng.*, vol. 12, no. 2, pp. 90–99, 2018.
- [13] C. Ge, "A LSTM and graph CNN combined network for community house price forecasting," in *Proc. 20th IEEE Int. Conf. Mobile Data Manage. (MDM)*, Jun. 2019, pp. 393–394.
- [14] S. Law, B. Paige, and C. Russell, "Take a look around: Using street view and satellite images to estimate house prices," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 5, pp. 1–19, Nov. 2019.
- [15] X. Fu, T. Jia, X. Zhang, S. Li, and Y. Zhang, "Do street-level scene perceptions affect housing prices in Chinese megacities? An analysis using open access datasets and deep learning," *PLoS ONE*, vol. 14, no. 5, May 2019, Art. no. e0217505.
- [16] Q. You, R. Pang, L. Cao, and J. Luo, "Image-based appraisal of real estate properties," *IEEE Trans. Multimedia*, vol. 19, no. 12, pp. 2751–2759, Dec. 2017.
- [17] O. Poursaeed, T. Matera, and S. Belongie, "Vision-based real estate price estimation," *Mach. Vis. Appl.*, vol. 29, no. 4, pp. 667–676, May 2018.
- [18] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1412–1421.
- [19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
- [20] O. Firat, K. Cho, and Y. Bengio, "Multi-way, multilingual neural machine translation with a shared attention mechanism," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 866–875.
- [21] P.-Y. Huang, F. Liu, S.-R. Shiang, J. Oh, and C. Dyer, "Attention-based multimodal neural machine translation," in *Proc. 1st Conf. Mach. Transl., Volume 2, Shared Task Papers*, 2016, pp. 639–645.
- [22] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 577–585.
- [23] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent NN: First results," 2014, *arXiv:1412.1602*. [Online]. Available: <https://arxiv.org/abs/1412.1602>
- [24] N. Moritz, T. Hori, and J. L. Roux, "Triggered attention for end-to-end speech recognition," in *Proc. ICASSP-IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 5666–5670.
- [25] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1240–1253, Dec. 2017.
- [26] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4651–4659.
- [27] X. Zhou, Y. Shen, Y. Zhu, and L. Huang, "Predicting multi-step citywide passenger demands using attention-based neural networks," in *Proc. 11th ACM Int. Conf. Web Search Data Mining*, Feb. 2018, pp. 736–744.
- [28] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based LSTM and semantic consistency," *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 2045–2055, Sep. 2017.
- [29] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 21–29.
- [30] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proc. IEEE*, vol. 103, no. 9, pp. 1449–1477, Sep. 2015.
- [31] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi, "Attention-based multimodal fusion for video description," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4193–4202.
- [32] J. Xu, T. Yao, Y. Zhang, and T. Mei, "Learning multimodal attention LSTM networks for video captioning," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 537–545.
- [33] Z. Yue, S. Ding, L. Zhao, Y. Zhang, Z. Cao, M. Tanveer, A. Jolfaei, and X. Zheng, "Privacy-preserving time series medical images analysis using a hybrid deep learning framework," *ACM Trans. Internet Technol.*, vol. 37, no. 4, pp. 1–22, May 2020.
- [34] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.



PEI-YING WANG received the B.S. degree in information management from National Central University, Taoyuan, Taiwan, in 2018, and the M.S. degree in information management from National Chiao Tung University, Hsinchu, Taiwan, in 2020. Her research interests include deep learning, artificial intelligence, and financial technology.



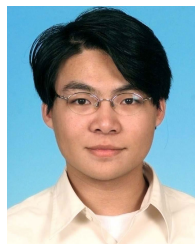
CHIAO-TING CHEN received the B.S. degree in information management and finance and the M.S. degree in information management from National Chiao Tung University, Hsinchu, Taiwan, in 2018 and 2020, respectively. She is currently pursuing the Ph.D. degree in computer science with National Yang Ming Chiao Tung University, Hsinchu. Her research interests include deep learning, artificial intelligence, knowledge graph, graph embedding, and financial technology.



TING-YUN WANG received the B.S. degree in management science from National Chiao Tung University, Hsinchu, Taiwan, in 2019, where she is currently pursuing the M.S. degree in information management. Her research interests include deep learning, artificial intelligence, and financial technology.



JAIN-WUN SU received the B.S. degree in information management from Yuan Ze University, Taoyuan, Taiwan, in 2019. He is currently pursuing the M.S. degree in information management with National Chiao Tung University, Hsinchu, Taiwan. His research interests include deep learning, artificial intelligence, and financial technology.



SZU-HAO HUANG (Member, IEEE) received the B.E. and Ph.D. degrees in computer science from National Tsing Hua University, Hsinchu, Taiwan, in 2001 and 2009, respectively. He is currently an Assistant Professor with the Department of Information Management and Finance and the Chief Director with the Financial Tehnology (Fin-Tech) Innovation Research Center, National Yang Ming Chiao Tung University, Hsinchu. He is also the Principal Investigator of the MOST Financial Technology Innovation Industrial-Academic Alliance and several cooperation projects with leading companies in Taiwan. He authored more than 50 papers published in the related international journals and conferences. His research interests include artificial intelligence, deep learning, recommender systems, computer vision, and financial technology.

...