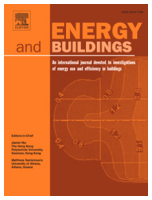
Energy & Buildings 338 (2025) 115723

Contents lists available at ScienceDirect

Energy & Buildings

journal homepage: www.elsevier.com/locate/enbuild



Learning from other cities: Transfer learning based multimodal residential energy prediction for cities with limited existing data

Yulan Sheng [a,*], Hadi Arbabi [a], Wil Oc Ward [b], Martin Mayfield [a]

[a] School of Mechanical, Aerospace and Civil Engineering, The University of Sheffield, Sir Frederick Mappin Building, Sheffield, S1 4DT, UK
[b] School of Computing and Mathematical Sciences, University of Leicester, University Road, Leicester, LE1 7RH, UK

ARTICLE INFO

Keywords:
Residential energy prediction
Transfer learning
Multimodal learning
EPC
Street view images

ABSTRACT

Reliable prediction of residential energy consumption is essential for informing energy efficiency policies and retrofit strategies. However, traditional data-driven approaches are often constrained by the availability and quality of data. This study presents a novel approach combining multimodal neural networks with a transfer learning framework, leveraging both tabular and visual data to enhance prediction accuracy and enable knowledge transfer from data-rich to data-poor regions. Case studies conducted in Barnsley, Doncaster, and Merthyr Tydfil demonstrated that the proposed approach outperforms traditional mono-modal models. The multimodal model improved prediction accuracy significantly, achieving a MAPE reduction from 1.15 (with only visual data) and 0.86 (with only tabular data) to 0.43 (with both visual and tabular data), while the inclusion of transfer learning offers further performance improvements in data-scarce regions, with up to 63.6 % error reduction. Explainable AI is utilised to validate the model's interpretability, confirming key features such as floor and wall insulation conditions as pivotal in energy consumption predictions. This integrated framework offers actionable insights for policymakers, facilitating data-driven decisions to enhance energy efficiency in diverse urban settings.

1. Introduction

Reliable prediction of operational energy consumption in residential buildings is crucial for guiding energy efficiency policies, supporting retrofit strategies, and contributing to broader sustainability goals. According to the UK Government, a reduction of at least 68 % of greenhouse gas (GHG) emissions at 1990 levels is required by 2030 [1]. Among all the energy users, the domestic sector is the second largest energy consumer, accounting for nearly 30 % of total energy consumption in the UK in 2022 [2]. Decarbonising the built environment is one of the key drivers to achieve the net-zero goals.

To support net-zero policies and address fuel poverty, various regulations and incentives, such as the Home Upgrade Grant [3], have been introduced to improve home energy efficiency. These initiatives primarily target residents at risk of fuel poverty or living in properties with an energy rating of D or below, through measures including insulation upgrades and implementing low-carbon heating technologies. However, there remains a need to strengthen the evidence base to identify the most effective retrofitting measures for different buildings and regions [4,5].

Traditional residential energy consumption estimation by data-driven models rely heavily on the availability of high-quality data, which is often unevenly distributed across regions. In many cases, data can be sparse or of poor quality, leading to inaccurate energy predictions and inefficient policy decisions. Recent advancements in deep learning offer promising solutions to these challenges, specifically, the multimodal neural networks and transfer learning. Although relatively new to building energy estimation, these approaches have demonstrated their ability to enhance modelling performance in other research domains, such as robotics and medical diagnostics [6–9]. Compared to models utilising a single data source, multimodal networks can integrate diverse data types, such as tabular and visual data, to provide a more comprehensive understanding of the factors influencing energy consumption. And transfer learning can further optimise the process by allowing models trained on data-rich region to be adapted for use in data-poor regions, thereby improving prediction accuracy across different spatial contexts.

This paper proposes a novel approach that combines these two techniques to enhance the prediction of annual residential energy consumption. We introduce the application of a deep multimodal neural network that leverages both tabular and visual data, integrating with a transfer learning element that facilitates knowledge transfer from data-rich to data-poor regions. The main research question is: **Can the reliability of residential energy consumption predictions be improved in regions with limited data availability through the integration of multimodal learning and transfer learning?**

* Corresponding address: School of Earth and Environment, University of Leeds, Leeds, LS2 9JT, UK.
E-mail address: y.sheng1@leeds.ac.uk (Y. Sheng).

https://doi.org/10.1016/j.enbuild.2025.115723
Received 13 December 2024; Received in revised form 18 March 2025; Accepted 6 April 2025
Available online 8 April 2025

Abbreviations

AP	Average Precision
BIM	Building Information Modelling
CNN	Convolutional Neural Network
CV	Coefficient of Variation
D_SBarnsley	The Barnsley source domain of the transfer learning model
D_SDoncaster	The Doncaster target domain of the transfer learning model
D_SMerthyr Tydfil	The Merthyr Tydfil target domain of the transfer learning model
D_TBarnsley	The Barnsley target domain of the transfer learning model
EPC(s)	Energy Performance Certificate(s)
GSV(s)	Google Street View(s)
LiDAR	Light Detection and Ranging
LLM	Large Language Models
MAPE	Mean Absolute Percentage Error
MARVEL	Multi-spectral Advanced Research Vehicle
MLP	Multi-layer Perceptron
NAS	Neural Architecture Search
OLS	Ordinary Least Squares
SVM	Support Vector Machine
SHAP	Shapley Additive Explanations
UPRN	Unique Property Reference Number
XAI	Explainable artificial intelligence

Symbols

R^2	Degree of determination
y	Ground truth of the dependent variable, $kWh/year$
$\hat{y}$	Predicted dependent variable $kWh/year$
$P(y\|x)$	Marginal probability distribution
$M!$	Number of coalitions
S	Coalition
$\|S\|$	Number of features in coalition S
ϕ	SHAP values

To validate the effectiveness of the proposed application of a multimodal and transfer learning approach, three scenarios are designed:

- **Scenario 1 Same city, different data source:** This scenario evaluates how well the proposed approach performs within a single city when using data from different collection means.
- **Scenario 2 Cities with similar building features, different data source:** This scenario evaluates how well the proposed approach can generalise predictions between cities with similar characteristics but different data acquisition methods.
- **Scenario 3 Cities with different building features, different data source:** This scenario evaluates how well the proposed approach transfers knowledge effectively across regions with varied architectural or geographical characteristics.

From these scenarios, this study develops and compares model performance in predicting yearly average energy consumption. It first assesses whether multimodal learning and transfer learning enhance prediction accuracy in regions with limited data availability. Then, an explainable AI tool is implemented to ensure that the model effectively extracts and transfers relevant information while identifying key features for residential energy estimation. The proposed approach is validated through case studies in three cities, demonstrating its effectiveness in improving prediction robustness and providing actionable insights for energy policy and retrofit strategies.

2. Related work

The residential energy consumption estimation has been intensively explored by three main approaches: physics-based, data-driven, and hybrid. Physics-based approaches typically rely on detailed information on buildings' thermal characteristics, for example, the thermal transmittance of the material used in building constructions. This data is then input into a physical model developed based on theories of heat transfer to estimate the properties' energy performance [10]. Benefiting from using only the physical characteristics of the property, this approach can be applied without knowledge of historical consumption data. On the other hand, the physics-based approach has large uncertainties in data processing, parameter assumptions and model settings and can be time-consuming. These limitations have restricted the widespread application of physics-based methods in large-scale studies. When detailed information and internal access are limited, or in large-scale studies, data-driven approaches are often adopted. The primary method of the data-driven approach is to develop machine learning models for estimation, based on historical energy consumption and building morphology. Hybrid approaches aim to minimise the limitations of both data-driven and physics-based methods by integrating them into different stages of a comprehensive framework [11]. This integration can offer significant advantages but also has notable limitations. One key challenge is the complexity of integrating the two models, knowledge of both data science and building physics is required. The lack of standardised frameworks further limited the approach from replication and generalisation. The hybrid approach also tends to be computationally expensive, as it requires detailed simulations being run alongside machine learning algorithms [12].

This research adopts a data-driven approach due to the lack of internal access to individual buildings and the focus is on conducting estimation at a city scale. Existing data-driven studies have explored various databases as indicators and compared the effectiveness of a wide range of algorithms. Table 1 provides a summary of selected existing studies in residential energy consumption estimation. Commonly used algorithms include decision trees [13–15], neural networks [13,16–18], k-nearest neighbours [13], and linear regressions [13,15]. For instance, [17] compared the effectiveness of nine machine learning models in estimating energy rating values using Dublin Energy Performance Certificates (EPCs) and concluded that deep learning algorithms performed best, with a Root Mean Square Error of 0.2. Similarly, [13] tested the performance of nine algorithms in predicting annual average energy consumption using a combination of building and meteorological data. The study included 285,000 residential buildings across ten cities in the UK and found that deep neural networks were the most efficient model, achieving a mean absolute error of 0.92. Studies, such as [16], utilised visual data, by dividing real estate images into individual patches as indicators for specific housing features. ResNet were used and achieved a classification accuracy of 62 % for predicting energy ratings.

Despite these promising results, the reliability of machine learning approaches faces two main limitations: trustworthiness and adaptability. Trustworthiness refers to the suitability of the data and model, and whether adequate performance has been achieved for the specific task. For energy consumption predictions, existing models often rely on single-modal networks, which are vulnerable to issues related to data noise and quality, and hence less trustworthy. For instance, the widely used data source in recent studies, the EPC, is believed to be problematic. Apart from the major drawbacks of EPC being a standardised representation of the property's energy usage, its accuracy is largely dependent on the individual assessor. A 'mystery shopper' study was carried out by [22] and [23]. Four assessors and an external organisation were asked to inspect the same 29 houses in the UK and then compare their notes and the resulting EPCs. Although the number of assessments is limited (because of the lack of qualified energy assessors [22]), the distribution of resulted ratings still provides insights into the significant inconsistencies in EPC calculation caused by the inspections. Notably,

Table 1
A selective summary of recent studies, including the study location, size of the input, algorithms used or found best, the housing features used and output. The existence of a check mark means the variable is used in the study. The variable is marked with an asterisk star if it is considered as the key housing feature in the corresponding energy analysis. The table is ordered according to the year of publication.

Author	[19]	[16]	[17]	[20]	[21]	[13]	[18]
Study location	New York, USA	Austria	Ireland, UK	Germany	UK	UK	Glasgow, UK
Input size	20,000	3865	850,000	25,000	All EPCs	5000	165,318
Algorithm	OLS	CNN	Neural Network	SVM	Logistic regression	DNN	CNN + MLP
Output	Energy intensity	EPC rating	EPC rating	Consumption	EPC rating	Consumption	EPC rating
Total floor area	✓		✓	✓*	✓	✓*	✓
Number of floors	✓				✓	✓	✓
Walls condition			✓	✓	✓	✓	✓
Windows condition			✓	✓	✓	✓	✓
Roof condition			✓	✓	✓	✓*	✓*
Floor condition	✓		✓		✓	✓	✓
Lighting condition			✓		✓	✓	✓
Fuel type			✓	✓	✓		✓
Main heat			✓	✓			✓*
Heating control							✓
Room Count					✓	✓*	✓
Year of construction	✓	✓	✓*	✓	✓*		✓
Building type			✓*	✓	✓	✓	✓
Location					✓		✓
Climate data						✓	
Images		✓					✓*

almost two-thirds of the assessed properties have had ratings varied across two EPC bands.

These limitations in mono-modal predictions have led to growing interest in deep multimodal learning, developing models that use multiple data inputs [24–26]. This approach involves integrating heterogeneous cues from different modalities, such as image (visual), text (word), and audio (sound), to provide more comprehensive knowledge for the given task. While this approach has seen applications in fields like face recognition, medical diagnosis, and self-driving systems [24,25], its use in building energy prediction remains limited. [18] examined the potential of multimodal approaches by combining Scottish EPC data with Google Street View (GSV) images to estimate energy efficiency ratings. The study, which examined 165,318 properties in Glasgow, found that including GSV facade images increased modelling accuracy from 79.7 % to 86.8 % compared to using only EPC data.

Although the street view image database has been largely enriched by the recent advances in technology, studies using street view images are prone to general challenges. Common obstacles include heterogeneous image quality, the presence of irrelevant objects, and variations in spatial coverage and update frequency [27]. Taking the images by driving through neighbourhoods also suggests that only the characteristics of the front facades of properties are considered in the model training, which largely neglected the features of the rear. These limitations suggest that, the same as tabular data, using street view images as the only input data may also lead to biased representations of the properties' energy performance.

Additional to statistical evaluations of model performance, the interpretability and explainability of machine learning models are critical in determining their trustworthiness and transparency. Black-box models have significant uncertainty regarding whether they genuinely understand the designed tasks. As models become increasingly complex, interpretability becomes more challenging. This difficulty in understanding complex algorithms emerges in the development of Explainable artificial intelligence (XAI). One of the popular XAI tools is SHapley Additive exPlanations (SHAP).

SHAP is developed by [28] based on the concept of cooperative game theory. The SHAP value represents the average marginal contribution of a feature across all possible feature coalitions [7,9,28]. Mathematically, SHAP values are computed as Eq. (1):

$$\text{SHAP values } \phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M-|S|-1)!}{M!} \left[f_x\left(S \cup \{i\} - f_x(S)\right)\right] \quad (1)$$

Where $M!$ is the number of ways to form a coalition, and $|S|$ is the number of features in the coalition S. The shapely value assumes the features join the coalition sequentially, which results in $(M-|S|-1)!$, representing the number of possible orderings for a feature to join after feature i. $\frac{|S|!(M-|S|-1)!}{M!}$ serves as the weighting factor for the marginal contribution of feature i to coalition S. A positive SHAP value indicates that the presence of a feature increases the model's predicted output, while a negative value suggests a decreasing effect. The mean absolute SHAP value (mean|SHAP values|) is commonly used as a measure of feature importance, providing a ranking of feature contributions.

Example study [18] used SHAP to examine the importance of building features in a multimodal prediction model. The study found that SHAP identified key pixels clustering around structural elements such as windows and doors, highlighting their significance in the energy prediction model. This demonstrates the utility of SHAP in enhancing model interpretability by pinpointing influential input features.

Adaptability, on the other hand, refers to the challenge in traditional machine learning where prediction data must have the same feature distribution as the training data, otherwise requiring a new model to be trained from scratch [29–31]. In many cases, data availability is limited, either due to low quality or the high cost of data collection. Transfer learning offers a solution that reduces the reliance on data by leveraging the knowledge gained from one task to help other relevant prediction tasks. For example, [30] used transfer learning to predict electricity consumption in a newly developed office building by learning energy behaviours from similar buildings in other cities, resulting in a 20 % average improvement in model performance. Similarly, [29] developed a transfer learning model to predict the energy demand for a building in the next 24 h, on the basis of a CNN model and transfer learning. The performance of this transfer learning model is evaluated by comparing it with a linear model, a CNN model and a model pre-trained without transferring the knowledge. The transfer learning model significantly increased the prediction accuracy by 20 %, 17 % and 30 % respectively.

Apart from the common limitations of black-box approaches, a key challenge in applying transfer learning to enhance prediction robustness is the possible mismatches in source and target domains. Significant discrepancies in domains may reduce the relevance of the transferred knowledge and lead to negative learning [32,33]. For example, features learned from the energy consumption patterns of one building type may not align well with those of another. In models involving visual data, discrepancies can also arise in fundamental image features, such as the number of channels, intensity, or orientation. Over-fitting during fine-tuning is another concern, especially when the target dataset is

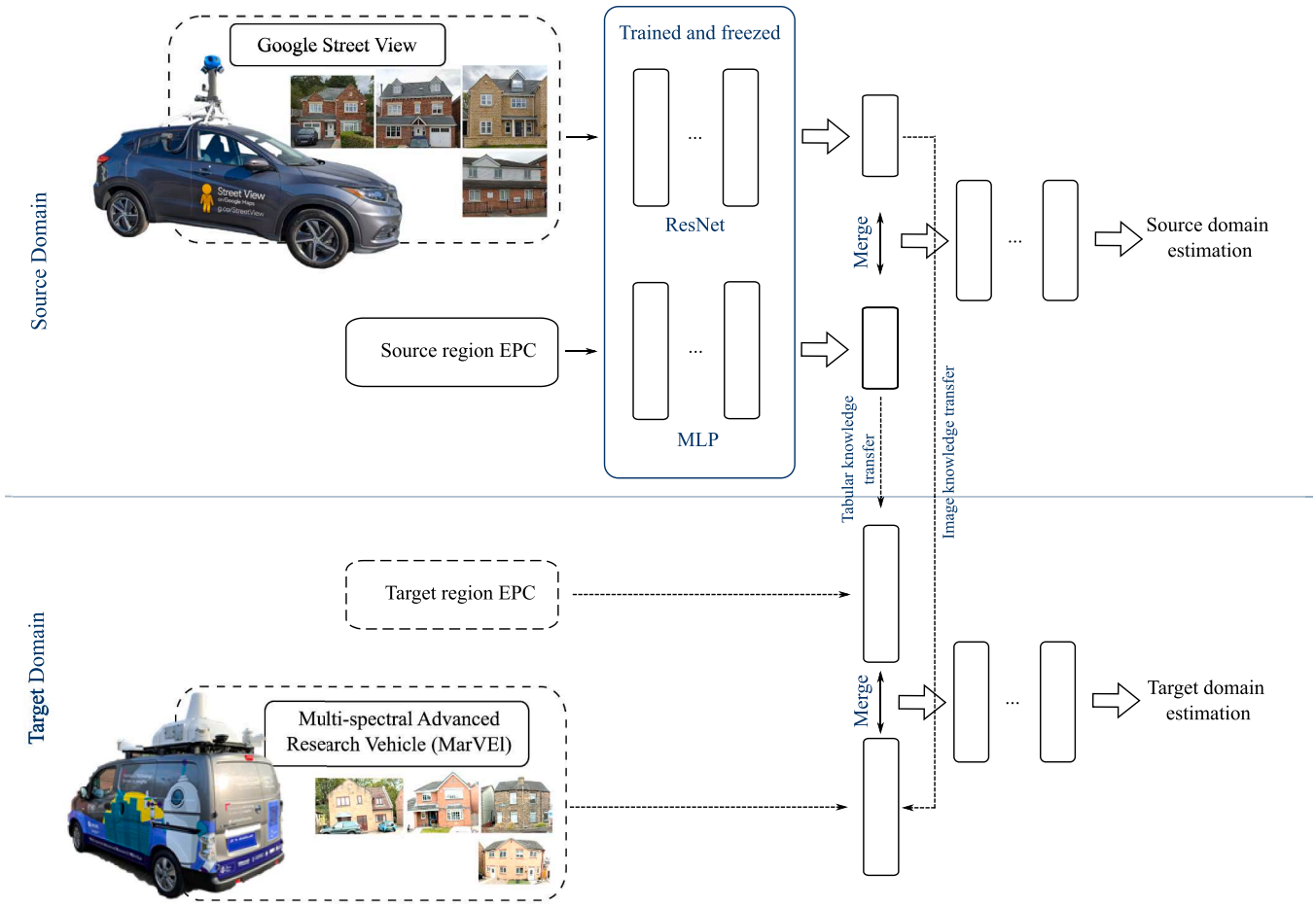


Fig. 1. Overall workflow. Knowledge is firstly learned from the source domain by training and freezing the shallow layers. Then transfers to the target domain before two modalities merge to perform the final prediction.

small, which may lead to poor generalisability. In addition, while there are studies that implemented transfer learning to predict energy usage, similar to other related studies, most of them rely on a single modality [29,30,33,34].

This study builds on these challenges by proposing a novel deep multimodal neural network incorporating a transfer learning component. Specifically, the methodology is designed to address the limitations of existing studies, where the reliability and transparency of predictions are constrained by their dependence on single-modality data and are further impacted by uneven data availability across regions. To evaluate the proposed approach, data were collected from three case study cities: Barnsley, Doncaster, and Merthyr Tydfil. The performance of the proposed framework is assessed by comparing it with conventional neural networks that rely on a single data source and lack transfer learning elements. Additionally, explainable AI techniques are employed to ensure that the model accurately identifies and transfers relevant information, thereby enhancing the interpretability of residential energy patterns and improving the overall robustness of energy estimation models.

3. Data and methodology

The technical framework employed in this study is illustrated in Fig. 1. The model is developed using AutoKeras, an automated Neural Architecture Search (NAS) tool that provides an efficient and robust approach for constructing deep learning algorithms [35].

The core structure of the workflow is a deep multimodal learning model, depicted in the first half of Fig. 1 as the "Source Domain." This multimodal framework integrates tabular and visual data, where tabular data is processed through a MLP and visual data through a ResNet-based CNN. Extracted features from both modalities are aligned using a common spatial reference, the Unique Property Reference Number (UPRN), to maintain consistency across data sources. The concatenated features are then utilised to predict annual residential energy consumption.

To enhance model generalisation in regions with limited data availability, the learned features from the lower layers of the multimodal model are frozen and transferred to the target domain. Fine-tuning is subsequently performed using available target domain data, ensuring adaptability to new urban contexts. The visual data for the target domain is sourced from the Multi-spectral Advanced Research Vehicle (MARVEL) [36].

A comprehensive discussion of the data sources and model development is provided in the following sections.

3.1. Data sources

The study utilises two primary data sources: EPCs and street-view images. EPCs provide detailed information about the physical characteristics of the buildings, and street view images offer additional contextual information capturing the exterior features.

Datasets for properties in three UK cities: Barnsley, Doncaster and Merthyr Tydfil, are collected for the designed scenarios, representing different data conditions commonly seen.

Table 2
List of data selected and extracted from the EPC, with brief description and example classes in categorical data. Detailed statistics and classes are included in Appendix.

No.	Variables	Description
1	Total floor area	Area of the building footprint (m^2)
2	Property type	Type of property (e.g. house)
3	Built form	Type of built-form (e.g. detached)
4	Number habitable rooms	Number of rooms in the property
5	Number heated rooms	Number of rooms can be heated in the property
6	Ageband	Construction age grouped in 12 bands (e.g. before 1900)
7	Roof description	Roof types and insulation conditions (e.g. pitched)
8	Walls description	Wall types and insulation conditions (e.g. filled cavity)
9	Floor description	Floor types and insulation conditions (e.g. solid, insulated)
10	Lighting description	Percentage of low energy lighting installed (%)
11	Main heat	Types of heatings used (e.g. Air source heat pump)
12	Energy consumption	Total energy consumption (kWh/year)

3.1.1. Tabular modality: energy performance certificates

This paper uses EPC as a tabular modality in multimodal network training. Although the issues with EPC have been intensively studied [22,23], it is one of the most comprehensive and publicly available databases with high spatial coverage for properties in the UK. Similar to the Energy Star score in the USA and Diagnostic de Performance Energétique in France [37], EPC is a legally required document in the UK for every property being sold or rented. The certificates are produced by qualified assessors, recording building information relating to the property's geographical location, building material and insulation conditions.

Using existing literature as reference, 11 out of 92 categories from the EPC are selected as inputs [13,16–21]. These features record the conditions of building elements and are directly associated with properties' thermal performance. Table 2 presents the selected features, short descriptions for each feature and an example class for categorical data.

Variables 1 to 6 are features describing the general characteristics of the buildings, and variables 7 to 11 provide more detailed descriptions of the conditions of specific building elements. Variable 12 is the mean annual energy consumption for each house, measured in kWh/year, which is used as the ground truth data to train the following energy prediction model.

Because the EPC records are usually created by multiple inspectors and may have also followed different versions of guidance, these tabular data were preprocessed to exclude inconsistencies and filter duplicated records. For example, if the entry is marked as 'INVALID!' or 'NO DATA', these entries are combined as 'unknown'. If the property address or reference number occurred multiple times, it means that the property is associated with multiple EPC records. These redundant EPCs were filtered based on when the record was created. The single latest-issued EPC is used as the data input. The classes in each categorical data were also reorganised. Similar descriptions in the categories are found and merged. For instance, 'some double glazing' and 'partial double glazing' used to describe the window insulation conditions are combined into one category.

3.1.2. Visual modality: street view images

To improve the reliability of the estimation, this work introduces the second modality into the model. Street view images can offer more information from the vertical illustration, including the appearance of the target properties, the allocation of and the ratios between different building elements, and also to some extent their conditions. Commonly, studies using street view images are prone to various general challenges as discussed [38]. Therefore, in this study, two databases were compared and combined to assess how these challenges may limit the accuracy of energy consumption estimations and how good data can help with the prediction with poor data, including a primary data collected by a van owned by the research group, named MARVEL, and a secondary data downloaded online through Google street view. Both the image database used in this paper were captured in a similar way: by driving a van through the neighbourhoods, capturing images alongside the road using multiple cameras and sensors equipped. In comparison to GSV images, the captured data by MARVEL has constraints in sample size, due to storage availability during each capture, and quality affected by the weather conditions. Linking the captured imageS with EPCS also suffered from significant data loss, further reducing the amount of image data available to use. The street view images are collected for three different cities, Barnsley, Doncaster and Merthyr Tydfil, a summary of these data is included in Table 3.

Table 3
A summary of the MARVEL image data used in this study for the three case study cities.

City	Sample size	Issues
Barnsley GSV	9050	
Barnsley capture	1547	Images overexposed
Doncaster capture	451	Limited sample size
Merthyr Tydfil capture	1345	Images affected by raindrops

The quality of the MARVEL captured image varies across regions. Example captures for the selected cities are presented in Fig. 2. The images captured in Barnsley are over-exposed (Fig. 2b), and some important features, especially the roof, become similar to the sky. The capture for Merthyr Tydfil was affected by the rainy weather (Fig. 2d). On the other hand, the weather conditions were considerably optimal during the capture in Doncaster (Fig. 2c), however, it primarily served as a test drive after the vehicle was initially set up which, as a result, the sample size are relatively small. In comparison, the downloaded images from Google provide more assurance of good quality and spatial coverage. The variations in quantity and quality of the captured data among the cities provide an interesting setting for this case study to showcase the potential of the transfer learning approach.

Unavoidably, street view images may contain visual information irrelevant to this study on building energy analysis. While existing studies such as [16] and [39] used scale-invariant feature transform to detect pixels of interest and produce individual small image patches with only specific building elements for prediction, we believe that using whole property images can allow the energy estimation algorithm better understand the global context of such housings. Therefore, to reduce the number of irrelevant contexts, an object detection algorithm, called YOLOv5, is applied to the extracted houses from the street view images. The YOLOv5 detects objects by dividing an image into a grid and then calculating the weights to help determine the possibility of whether the detected pixels belong to a house feature as a regression problem [40]. A custom YOLOv5 model is trained specifically for this study using over 800 manually labelled GSV images with bounding boxes, and achieved an average precision (AP) of around 0.8, suggesting the trained model successfully detected where houses are located in the street view images.

(a) Barnsley (G) (b) Barnsley (M) (c) Doncaster (M) (d) Merthyr Tydfil (M)

Fig. 2. Example street view images captured by two different sources used in this study. The quality of the image varies across regions, mostly affected by weather conditions.

If multiple houses are detected in the same image, the largest house detected is used for the following prediction.

The images for the detected houses are further processed to remove the sky using watershed segmentation. As skies usually have brighter appearances than properties, this step avoids the developed model identifies skies as the key feature for estimation. All the processed images are stretched and resized when necessary and were saved with the UPRN references to match with their corresponding EPC records to facilitate the following multimodal learning.

3.2. Multimodal network based architecture

To capture the diverse factors influencing residential energy consumption, this study developed a deep multimodal neural network. The model architecture consists of two main branches:

1. **Tabular Data Processing Branch:**
   This branch processes the EPC data, which includes variables such as property type, insulation conditions, and recorded energy consumption. The data is passed through a MLP structure, which is constructed with a series of fully connected layers that transform the input features into higher-dimensional representations, capturing complex relationships between different building attributes.
2. **Visual Data Processing Branch:**
   The visual data, provided in the form of street view images, is processed using CNN based architectures. The CNN extracts features from the images, which are relevant for energy consumption estimation, such as roof shape and wall materials.

After processing by their respective branches, the outputs from both data streams are merged into a single value, and then passed through additional dense layers to generate the final energy consumption prediction.

3.2.1. Model performance: mono-modal vs multimodal

A comparison is conducted between mono-modal and multimodal architectures, to evaluate the effectiveness of including a second source of data in estimation. Common algorithms used in existing literature is selected for evaluation, CNN and ResNet are employed for image data and MLP is selected for the tabular data [16,18].

The performance of the networks is evaluated using R^2 and MAPE. The evaluation results were used for comparisons between different algorithms and as evidence to select the best algorithms for multimodal network construction. R^2 is the coefficient of determination measuring the degree of fitness. Although the value range for R^2 is usually between 0 and 1, where close to 1 suggests the model is a good fit for the data inputs, it is possible to be arbitrarily negative, presenting worse performance. A negative R^2 suggests the difference between $\hat{y}$ and y is significantly large, indicating the model is not explaining better than just a single line plotted on the mean value. MAPE calculates the average percentage difference between the predicted values and the expected ground truth values. It is usually presented in a percentage format, ranging from 0 % to 100 %, but can be over 1 suggesting a higher error rate.

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

$$\mathrm{MAPE}(y,\hat{y}) = \frac{1}{n}\sum\left(\frac{|y_i - \hat{y}_i|}{y_i}\right) \times 100 \tag{3}$$

According to the evaluation results, the best architecture is used to develop a multimodal network, by adding a concatenate layer after each stream of data being processed and computed. Four models are trained in this case and assessed whether the multimodal network is able to produce more accurate prediction results. This evaluation ensures the network is able to capture hierarchical patterns and representations from the inputs that are meaningful to energy estimation, which further ensures useful information is identified and leveraged in the next stage of the designed transfer learning model.

3.3. Transfer learning for cross-regional adaptation

The next stage is to address the challenges posed by limited data availability in certain regions, a transfer learning component was incorporated into the multi-modal model. The idea of transfer learning mainly revolves around three elements, the domain, D, the prediction task T, and the marginal probability distribution $P(y|x)$. Transfer learning considers the base model, or in this case, the region where comprehensive housing data is available, as the source domain D_S. And the region where less data is available, as the target domain D_T.

The multimodal transfer learning model proposed in this work is developed and optimised through the utilisation of AutoKeras. AutoKeras is an open-source AutoML built on top of TensorFlow and Keras. It automates the process of model selection, hyperparameter tuning, and training optimisation. It employs neural architecture search and Bayesian optimisation techniques to ensure an efficient and optimal model configuration [41,42]. In this study, the AutoModel package was employed to develop the multimodal transfer learning model.

The integration of transfer learning and multimodal learning begins with training the multimodal network on a source domain comprising regions with relatively better-quality and better-coverage data, then fine-tuning on a target domain with poor data availability. The source domain uses the multimodal network as the base architecture. The best models found following the comparison study discussed in Section 3.2 are specified when constructing the AutoModel, by using a DenseBlock for tabular data and a ResNetBlock for visual data.

Once the model is trained on the source domain, the lower layers of the network, responsible for feature extraction, are frozen (by setting *trainable* into *False*), to preserve the generalized features they have learned. The upper layers, which handle the final prediction, are then

Table 4
Results of MMD between each domain of the case study cities.

Data	D_T Barnsley	D_T Doncaster	D_T Merthyr Tydfil
Tabular		0.136	0.096
Visual	0.017	0.045	0.011
Multimodal	0.007	0.024	0.009

fine-tuned using AutoKeras with the target domains. The learning rate for the trainable layers is set as ten times higher than the untrainable layers, so the trainable layers become task-specific, while the untrainable layers can keep useful representations transferable through tasks. The performance of the transfer learning model is validated across the target regions, to assess how well the model can adapt and maintain accuracy despite the reduced data quality and quantity in these areas.

3.4. Interpretability and explainability

The models developed in each stage were further assessed using SHAP to validate their capability in identifying and transferring key information from both modalities for energy estimation. At the same time, the feature importance ranking ensured the model's predictions were interpretable and actionable.

This interpretability is crucial for policymakers and energy planners who need to justify retrofit decisions and prioritise interventions. SHAP values provide a way to quantify the contribution of each input feature to the model's predictions, making it easier to understand which factors are most influential in determining energy consumption. By identifying the most important features, the model may provide guidance to improve energy efficiency in residential buildings, in relation to which house and what element in house should be prioritise for retrofitting.

4. Case study in three cities

4.1. Overview

To explore the three designed scenarios: *Scenario 1: Same city, different data source*, *Scenario 2: Cities with similar building features, different data source*, and *Scenario 3: Cities with different building features, different data source*, three cities were studied as target regions: Barnsley and Doncaster, England, and Merthyr Tydfil, Wales.

Geographically, Barnsley and Doncaster are neighbouring cities, whereas Merthyr Tydfil is a town in Wales which located further away. According to the first law of Geography: 'Near things are more related than distant things' [43], which suggests Barnsley and Doncaster may have more similarities.

To quantify the discrepancies statistically, the Maximum Mean Discrepancy (MMD) was computed between each target domain and the source domain. For tabular data, since D_TBarnsley is a subset of D_SBarnsley, their distributions are nearly identical, and thus, D_TBarnsley was excluded from the comparison for tabular data. As presented in Table 4, all the MMDs are relatively low, indicating overall similarity among the domains. Larger discrepancies are found between D_SBarnsley and D_TDoncaster across all the data types. Where the greatest discrepancies are found in EPC data, with an MMD of 0.136, suggesting a moderate level of difference. Notably, the discrepancy between these two domains decreased when two modalities were incorporated. This preliminary analysis confirms that the datasets for the case study cities share common characteristics, supporting the feasibility of applying a transfer learning approach.

4.2. Source domain: the deep multimodal network

The source domain is trained with data from residential properties in Barnsley. A total of 13,384 EPC records were downloaded at the time of the study, with over 13 % of the properties found to have multiple entries. Therefore, filtering and reorganizing the data is essential. The overall statistics for the EPC data used can be found in Appendix A. Among these properties, approximately 72.9 % are houses (H), 20.4 % are flats (F), 6.1 % are bungalows (B), and the remaining 0.7 % are maisonette (M). The average energy consumption amounted to approximately 15,214 kWh per year which is lower than the national average. Some extreme cases were observed where recorded energy consumption exceeded 1000 kWh/m^2 per year, but no clear associations with specific housing characteristics could be identified for these extreme consumption levels, therefore, are excluded in this study.

Table 5
Model inputs, structure and evaluation results of all deep learning models developed in this case study, using either single or multiple modalities.

Model	Algorithm	Input data	R^2	MAPE
1	CNN	Image	0.76	1.73
2	RestNet	Image	0.84	1.15
3	MLP	Tabular	0.90	0.86
4	**Multimodal MLP + ResNet**	**Tabular & Image**	**0.97**	**0.43**

4.2.1. Comparison between mono-modality and multi-modality

Following the proposed methodology, four models were developed to predict the yearly average energy consumption of selected residential properties in the case study cities. Table 5 presents their evaluation results. Models 1 and 2 were trained using GSV images, Model 3 used EPC data, and Model 4 was a multimodal network incorporating both EPCs and GSVs.

All three models built on a single modality achieved an R^2 score over 0.7, which validated that these common algorithms found in existing literature are suitable for the purpose. Comparing the two models trained with GSV images, RestNet achieved a better performance than the model using CNN. Model 3 achieved the best MAPE results among the three models, which suggests that, when only using a single modality for residential building energy estimation, tabular modality may provide more effective indications regarding the underlying patterns of the relationship between housing elements and energy consumption, and thereby result in better predictions.

Based on the evaluation results, the RestNet and MLP were selected to build the multimodal network. By incorporating both sources of data, the model performance has been significantly improved. The MAPE value has dropped from over 100 % to 43 % with an R^2 value of 0.97, suggesting the multimodal network is able to provide a significantly better understanding of the property energy performance using both modalities.

4.3. Source domain feature importance

The impacts of incorporating multiple streams of data are also reflected in the changes in features' relative importance to estimation. SHAP was used to provide explanations on how the inputs contribute to the final prediction of the designed network. At the same time to examine whether the network effectively captures hierarchical patterns and representations relevant to energy estimation, ensuring the transfer of meaningful information.

A random selection of images is presented in Fig. 3 to demonstrate the SHAP explanation results. The first column in the figure is the original image input, the second column is the result identified by the model built by using only images, and the third column is the result of the model built with multiple modalities. Darker red suggests higher SHAP values, which means the pixels have higher positive impacts on the model performance, while darker blue suggests the opposite. Comparing the images in the same row shows the changes in key pixels identified by the respective model. It can be seen that fewer pixels were detected by the mono-modal network than by the multi-modal network,

Fig. 3. The changes in key features captured in images are presented. The columns, in order, represent the original image input, the key pixels identified by the model using only image data, and the key pixels identified by the model incorporating multiple modalities.

suggesting that it has a relatively weaker capability in extracting informative features from the facade. In the mono-modal network, pixels surrounding the edges of the property, especially the roof of the property, tend to have a higher contribution to estimating energy consumption. In comparison, after incorporating EPC data, the multimodal network is able to extract more useful features from the building facade. However, the key pixels identified by the multimodal network relatively show fewer clusters, and no clear evidence can be seen whether certain building features contribute more to the model prediction. This means that it is difficult to identify key features and design respective retrofit scenarios, emphasising the important role of the tabular modal included in the proposed network.

The changes in tabular modality, and the relative importance of the EPC features, are presented in Fig. 4. For the single modality model, the total floor area is identified as the most important feature for predicting the operational energy consumption of properties in Barnsley, followed by the properties' age band and the conditions of the wall. After adding images as another source of data, the rank changed. The 'floor descriptions' becomes the dominant feature, followed by the conditions of walls and roofs. Total floor area and year of construction dropped to rank seventh and eighth respectively. This change may indicate that, after integrating visual modality, the network may have found that 'Floor Description' carries more information for the predictions, which might not have been apparent in the purely tabular setup.

Table 6
Results of the model training for three different case study cities, without and with using the transfer learning approach proposed in this study.

Case study cities	Without transfer		With transfer		
	R^2	MAPE	R^2	MAPE	Improvement in MAPE
Barnsley	0.28	7.12	0.70	2.92	59 %
Doncaster	-41.09	1.29	0.42	0.58	55 %
Merthyr Tydfil	-0.39	1.84	0.83	0.67	64 %

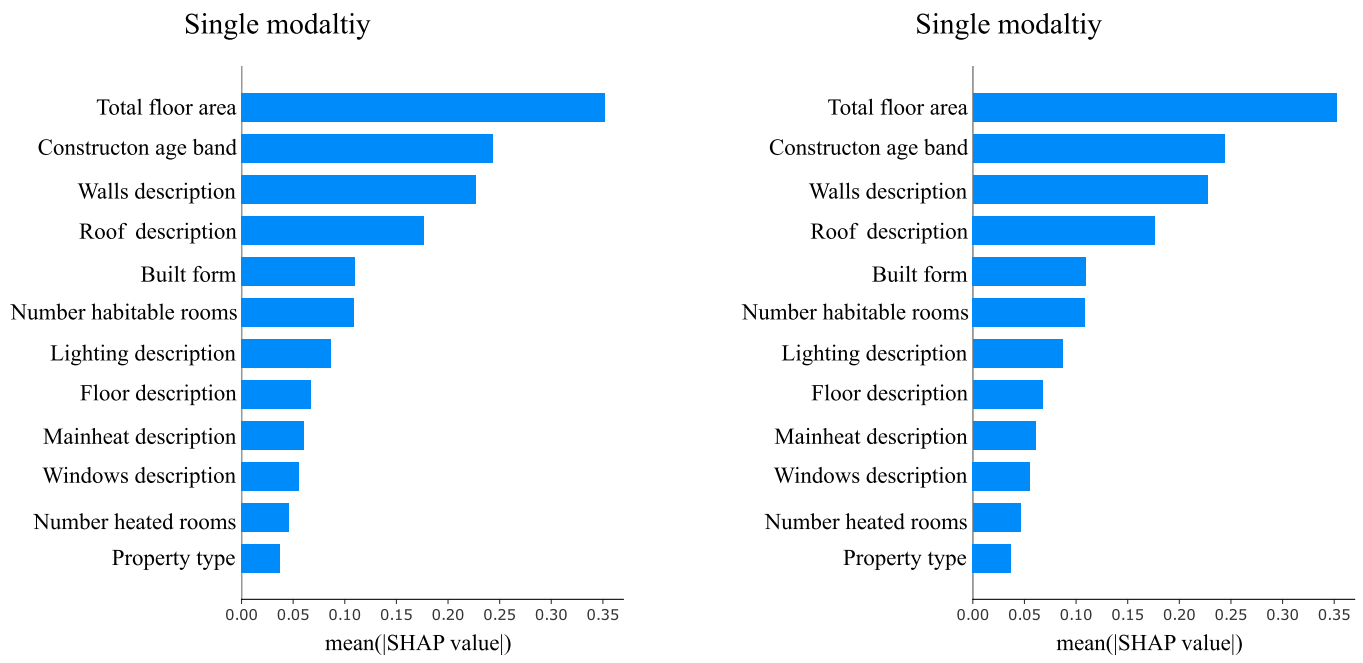
4.4. Target domain: the application of transfer learning

The evaluation results of the multimodal network have proven its capability of achieving trustworthy energy estimation for the assessed residential properties in Barnsley. Following the proposed methodology, this paper continues to explore the application of transfer learning in assisting prediction for regions with limited data availability. Two models were trained for each case study city, and their evaluation results are presented in Table 6. Two sets of configurations were tested to compare model performance with and without the transfer learning component in predicting annual energy consumption. Models without transfer learning were trained and fine-tuned solely on the source domain and tested on the target domain, without adapting to the target domain's data. In contrast, models with transfer learning utilised knowledge from the source domain by pre-training on the source data and then fine-tuning on both the source and target domains before testing on the target domain. This approach establishes a baseline for assessing the added value of transfer learning. Notably, the lower quality of MARVEL data resulted in reduced evaluation metrics compared to Table 5 where only GSVs are used as the visual data input. Graphic representations of the model prediction results are included in Appendix B.

Despite the multimodal network trained in Section 4.2 has demonstrated its capability of providing an accurate estimation of the operational energy performance for properties in Barnsley using EPCs and GSVs. However, without transfer learning, the models showed relatively low accuracy and fitness when applied to other target domains.

As expected, since the base model was initially trained using GSV and EPC from Barnsley, the model without integrating transfer learning component returned the best model fitness for D_TBarnsley, with the highest R^2 score of 0.28. For D_TDoncaster and D_TMerthyr Tydfil, the R^2 scores were significantly lower, with negative values indicating poor model performance and substantial errors between the predicted $\hat{y}$ and actual y values. These results suggest that the model trained solely on D_SBarnsley was not suitable for energy estimation in other cities. The differences in urban morphology and building characteristics, and also the differences between MARVEL-captured and GSVs are all likely to contribute to the poor transferability.

Incorporating the transfer learning component led to substantial improvements in model performance across the three case study cities. For Barnsley, the R^2 increased to 0.70, and the prediction error, MAPE, decreased from 7.12 to 2.92, representing a 59 % improvement in accuracy. For Doncaster, the R^2 improved from -41.09 to 0.42, with a corresponding reduction in MAPE from 1.29 to 0.58, yielding a 55 % improvement. For Merthyr Tydfil, the R^2 increased from -0.39 to 0.83,



(a) Single modality (b) Multi-modality

Fig. 4. Changes in the rank of relative feature importance of tabular inputs to the model. The upper bar chart is mono-modality, and the bottom chart is multi-modalities.

and the MAPE decreased from 1.84 to 0.67, representing a 64 % enhancement in model performance.

Larger improvements in prediction accuracy are observed for D_TMerthyr Tydfil compared with the other domains. This finding may be associated with the similarities between the domain feature distributions. As previously discussed in Table 4, the multimodal data of D_TMerthyr Tydfil exhibits a lower MMD with D_SBarnsley than with D_TDoncaster.

Overall, these results answer the research question and highlight the efficacy of transfer learning in improving model fitness and accuracy across diverse geographical contexts. The observed improvements underscore its potential for enhancing energy estimation in data-scarce regions.

4.5. Target domain feature importance

Although only limited data for the target domain was used to train the prediction model, the relative feature importance rank still offers valuable insights for these examined properties. The ranking is presented in Fig. 5.

In Barnsley, as illustrated in Fig. 5a, slightly different ranking results were shown comparing to the results in Fig. 4. The conditions of the building's structural elements, floor and roof, remain the dominant features for estimation. Surprisingly, the proportion of energy-saving lighting being installed is ranked third in the transfer learning based prediction. This change could be attributed to various factors. We can see from the statistics provided in the Appendix, that compared with the entire database used for training the base network, Table A.8, the reduced samples when acting as the target domain, Table A.17 have a larger value of coefficient of variation (CV) (69 %, the entire database has a CV of 55 %). This larger CV means that the properties in $D_T\,Barnsley$ have higher variability in their lighting conditions, which may be the reason for the increased importance being identified by the transfer learning model.

For the Doncaster domain, Fig. 5b, the conditions of the walls play a dominant role in the estimation, followed by property type and roof conditions. It is worth noting that, the EPCs used to tune Doncaster's transfer learning model only comprise a very small number of distinct classes for roof conditions, but it resulted in a high rank of importance. There are two possible interpretations for this rather contradictory finding. One is that, these classes of roof conditions have a pronounced effect on energy consumption estimates for Doncaster, so the model found higher importance for this feature. Another possibility is that, because the sample size for Doncaster is small, the data is less able to alter the 'knowledge' the base network learned from $D_S\,Barnsley$. As the roof is considered as the third important feature for estimating $D_S\,Barnsley$, the transfer learning model, before training with any target domain, may have naturally assigned a higher importance to the roof to start with.

The Merthyr Tydfil domain also exhibits different ranking results, as shown in Fig. 5c, where the conditions of walls, roof and year of construction are found as the most important features. The last four features, total floor area, main heating used, type of window and built form were found to be not important in this prediction. The reason is that, as statistics shown in Table A.34, all properties in this subset of the dataset contain quite similar classes. The total floor area for this subset of properties only has a CV of 0.13, suggesting a small variation around the average size. All of the properties examined are using 'Boiler', and 'Double glazing' windows.

5. Discussion

This research underscored the potential and benefits of integrating multimodal learning and transfer learning for residential energy consumption predictions. The significant performance gains observed across different regions highlight the value of combining diverse data sources and transferring knowledge across domains, especially for regions with limited available data sources.

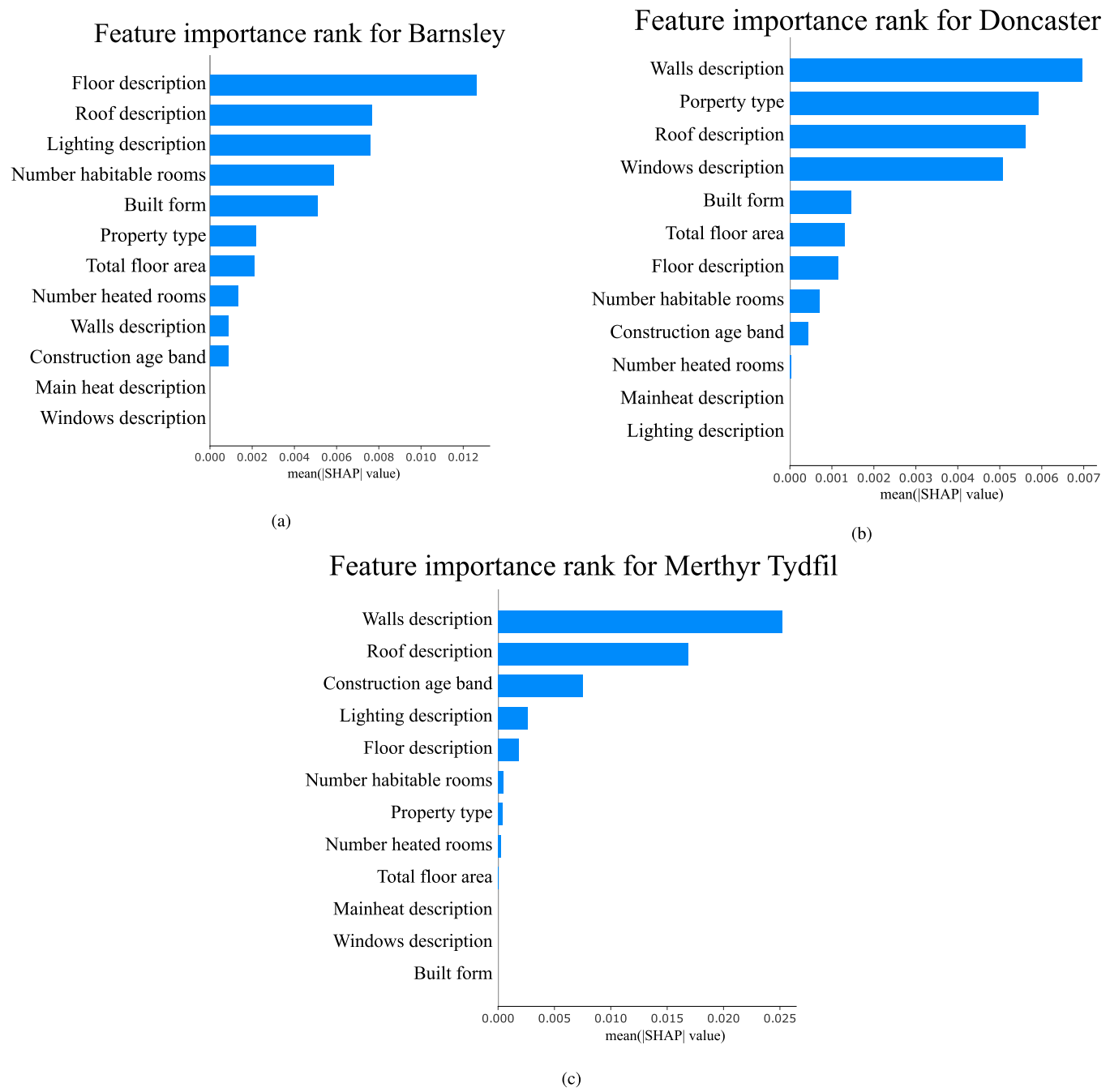


Fig. 5. The feature importance ranks for the three target domains, computed by the transfer learning models. (a) Feature importance rank for D_SBarnsley. (b)Feature importance rank for D_SDoncaster. (c)Feature importance rank for D_SMerthyr Tydfil.

The study first developed a deep multimodal network. The evaluation of multimodal networks against the single modal networks demonstrated significant improvements in prediction accuracy and fitness to the dataset. The reduction in MAPE to 0.43 and the high R^2 value of 0.97 underscore the robustness of multimodal models in capturing key energy determinants. These findings align with prior research emphasizing the limitations of relying solely on single data source, in particular the EPC data, due to inconsistencies and assessor biases [22,44], highlighting the added value of incorporating additional modalities in improving model reliability.

The inclusion of the transfer learning component demonstrated its effectiveness in adapting the model across regions with limited data. By leveraging knowledge from data-rich areas, the model significantly improved the prediction accuracy for data-poor environments, making it a robust tool for energy planners and policymakers. The most significant improvement was found for D_TMerthry Tydfil, the MAPE was improved by 64 % after employing the transfer learning approach.

In addition to achieving a more robust estimation, another key contribution of this paper is the exploration of key building determinants for consumption by using SHAP to enhance interpretability. Table 7 presented a summary of all the feature importance rank concluded in this work.

The first feature rank only included EPC records in the model training, while the rankings 2 to 5 were produced using both EPCs and images. The *roof descriptions*, which describe the type and insulation conditions of the roof, appeared the most number of times in the top 3 features among all the models developed. For ranking sorely built on EPCs, the *size of the property* is identified as the key estimator for Barnsley. Such

Table 7
All feature importance ranks produced by the models in this paper.

No.	Region	Rank 1st	Rank 2nd	Rank 3rd
1	Barnsley (mono-modal)	Total floor area	Age band	Walls description
2	Barnsley (multimodal)	Floor description	Walls description	Roof description
3	Barnsley (transfer)	Floor description	Roof description	Lighting description
4	Doncaster	Walls description	Property type	Roof description
5	Merthyr Tydfil	Walls description	Roof description	Age band

a finding is consistent with the inherent relationship between dwelling size and its energy consumption, as larger properties naturally require a larger number of energy-intensive appliances, such as heating facilities, to maintain residents' demand for comfortable living. It's contribution becomes less significant when another modality was introduced. This variations suggests that a multimodal model is able to learn more complex interactions and relationships between features across the modalities, which may lead to the discovery of hidden patterns that are not apparent when only individual modality is considered.

The models for Barnsley agreed that *floor description*, which describes the material and insulation conditions of properties' floor, is the most important feature for energy estimation for propoties in Barnsley. And models for Doncaster and Merthyr Tydfil both consider *walls descriptions*, which describes the material and insulation conditions for walls, are the key feature in their estimations.

These findings align with previous studies and the prioritisation of home upgrading measures implemented by the UK government. Currently, most home renovation projects focus on loft improvements, cavity wall insulation, and solid wall insulation, which corresponds with our feature importance analysis, where *roof description* and *walls description* emerge as key determinants of residential energy consumption. Beyond informing retrofitting priorities, these findings also provide insights into sensor placement and data collection strategies to strengthen the evidence base for future energy assessments [4,5]. For example, given the high importance of wall and roof conditions, deploying thermal imaging sensors, LiDAR scans, or street-view-based facade analysis could enhance data collection in energy audits. These approaches would improve the accuracy and granularity of energy performance assessments, supporting more data-driven decision-making in building energy policy and retrofit planning.

The differences between the feature importance ranks 2 to 5 approved that the proposed transfer learning model is able to derive patterns specific to the target domains and adapt the model accordingly so it is more suitable and robust for the target regions. This finding further emphasized that, the model used and the retrofit strategies should be regionally tailored rather than adopting a one-size-fits-all approach.

Compared to existing literature, this study extends the current understanding of residential energy prediction by providing empirical results on the effectiveness of transfer learning in bridging data gaps. While previous research has explored machine learning applications for energy efficiency, as discussed in Section 2 and the examples listed in Table 1, few studies have systematically examined the role of transfer learning in this domain. This work demonstrates that incorporating transfer learning not only improves model accuracy but also enhances its adaptability to diverse urban settings, paving the way for more scalable and equitable energy modelling solutions.

These insights have significant policy implications. The ability to accurately predict energy consumption with minimal data opens new opportunities for targeted energy interventions, particularly in cities with limited monitoring capabilities. Policymakers can leverage these models to identify high-risk properties, prioritise optimal retrofits, and develop cost-effective decarbonisation strategies. Furthermore, by integrating multimodal inputs, this approach reduces reliance on subjective EPC assessments, offering a more objective and data-driven foundation for energy policy decisions.

5.1. Limitations and potential future work

One of the primary assumptions this work based on is that the features recorded in the EPCs are accurate indications of the property. However, substantial studies have highlighted errors in EPC records [44,45]. In the proposed framework, this limitation is addressed by incorporating a second visual modality through street view images. These images provide supplementary information on building characteristics, such as window types, façade materials, and roof conditions, which can compensate for inaccuracies in EPC data. Nevertheless, EPCs still play a key role in model training and feature importance ranking.

Due to limited accessibility to individual properties, the main limitation of the image data used is that street view images usually only show the front exterior surface of the property. This might cause some key characteristics to be neglected by the model. For example, for properties with a similar front facade, properties with floor-to-ceiling windows at the back, may exhibit a large difference in energy consumption compared with properties with only solid walls.

While the official authorities are continuously updating the methodology for creating EPC, and new technologies are enriching the collection of street view images, future studies may include in-situ measurement of target properties to calibrate the entries in EPCs, more photos from multiple angles for the properties or create their own housing database. The recent advances in point-cloud based 3D building reconstruction and digital twins hold great potential to further improve the model accuracy. By using laser scanning or photogrammetry, it is possible to develop as-built BIMs that represent detailed building geometry with a high level of precision [46,47]. One of the example databases to be explored is the 3D city model derived from airborne LiDAR point clouds for Glasgow [48]. Following the framework this study proposed, using a calibrated or new database, future studies can contribute to a more holistic understanding of the operational energy performance of the existing housing stock, the estimation can be more accurate and trustworthy, the feature rankings and partial dependence can cover more features of interests, together leading to a thorough knowledge and evidence base for further explorations on retrofit potentials.

Although significant improvement in prediction performance was seen after incorporating transfer learning, the proposed multimodal transfer learning network did not achieve the highest accuracy observed when the model was trained directly on target city data ($R^2 = 0.70$ vs. 0.97).

This discrepancy may be a result of the inherent limitations of transfer learning in cross-regional energy estimation. Although we have examined the MMD metric before application to ensure domain similarity, the impacts on energy consumption patterns by the variations in urban morphology, building typologies, and socio-economic factors remain, making it difficult for the model to fully capture the localised energy dynamics of a new region, resulting in lower predictive accuracy compared to a model trained exclusively on target domain data. Furthermore, recent studies have argued that the network designed by NAS may not always outperform human-designed models [49]. While this paper lays the foundation of the integration between multimodal and transfer learning for residential energy consumption prediction, apart from collecting more high-quality data, future research may collaborate with experts in computer science and deep learning to explore alternative

network architectures and evaluate their performance against NAS-derived models.

In addition, while data-driven approaches provide a powerful means of estimating residential energy consumption, actual energy usage is highly dependent on occupants' energy use behaviours. Occupant behaviours, such as daily activities, thermal comfort preferences, and interactions with HVAC systems, are critical factors influencing energy consumption. Although the ground truth data used in data-driven approaches, usually historical consumption, implicitly captures some aspects of these behaviours, the gap between estimation and actual meter reading, as the discussed example research by Jenkins et al. [22] highlights the need to integrate dynamic data, for which the multimodal methodology proposed in this research lays a strong foundation. Recent advancements in Large Language Models (LLMs) present a promising opportunity to enhance predictive frameworks by processing vast datasets, such as interview transcripts, to extract nuanced behavioural insights and incorporate them into energy consumption models, thereby capturing complex patterns. Future research could explore a multimodal-LLM-transfer learning approach to improve energy estimation and policymaking, where multimodal data and LLMs provide contextual information on consumption patterns to enhance prediction reliability, and transfer learning ensures adaptability across different regions.

6. Conclusion

While there is no single solution to mitigate the climate crisis, decarbonising the existing housing stock, particularly in the residential sector, is a critical step. Effective decarbonisation requires a reliable and adaptable approach to understand current operational energy usage to guide the optimal retrofit measures. Although machine learning is a widely developed field, concerns remain, especially regarding the trustworthiness and adaptability of the existing methodologies. In response to these concerns, the idea of applying multimodal learning and transfer learning to operational energy estimation is introduced.

To achieve this, this work investigated the application of multimodal and transfer learning as an innovative approach to improve the reliability and adaptability of residential energy consumption estimation modelling. The effectiveness of the proposed approach was examined through three designed scenarios, *Scenario 1: Same city, different data source*, *Scenario 2: Cities with similar building features, different data source*, and *Scenario 3: Cities with different building features, different data source*. Using data from three case study cities, this paper evaluates the proposed application through statistical metrics and explainable AI tools.

Comparison studies were conducted to systematically evaluate the benefits of employing deep multimodal networks and transfer learning for energy estimation from three case study cities. By leveraging both tabular (EPC) and visual (street-view images) data, the deep multimodal neural network significantly outperformed conventional mono-modal models. The multimodal approach reduced the MAPE from 1.15 (image-only model) and 0.86 (tabular-only model) to 0.43, achieving an R^2 of 0.97, which demonstrates a substantial improvement in predictive accuracy. Furthermore, the incorporation of transfer learning proved highly effective in adapting models to data-scarce regions. Without transfer learning, the model performed poorly on target domains, with R^2 values as low as -41.09 and MAPE exceeding 1.84 in some cases. The transfer learning component improved model performance significantly across all tested regions, achieving a 59 % reduction in MAPE for Barnsley, 55 % for Doncaster, and 64 % for Merthyr Tydfil. The results demonstrated large improvements compared with the traditional modelling approach, underlining the merit of the proposed methodology.

The outcomes also offered insights into prospective retrofit prioritizations for the areas of investigation, through the use of explainable AI SHAP. The most important features identified in each study city can be used to design the potential retrofit measures for implementation. For example, for properties in Barnsley, possible retrofit options can be 1) improving floor conditions, 2) improving wall conditions and 3) improving both floor and walls. By integrating potential retrofit costs with the partial dependence values, it will be possible to identify the most cost-effective retrofit options for the target regions.

Ultimately, the proposed multimodal and transfer learning framework presents a robust and scalable approach to supporting data-driven energy policy and targeted retrofit strategies, contributing to the broader goal of reducing carbon emissions in the residential sector.

Additional Information

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Yulan Sheng: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization; **Hadi Arbabi:** Writing – review & editing, Supervision, Methodology, Conceptualization; **Wil Oc Ward:** Supervision, Data curation; **Martin Mayfield:** Writing – review & editing, Supervision, Funding acquisition

Acknowledgements

YS was supported by the University of Sheffield under the University Energy Flagship Institute Scholarship and the PGR Publication Scholarships. HA, WOCW and MM were supported by the EPSRC Active Building Centre [EP/V012053/1]. HA and WOCW were additionally supported by Towards Turing 2.0 under EPSRC, United Kingdom [EP/W037211/1] and The Alan Turing Institute, United Kingdom.

Appendix A. Statistics of the data used

A.1. Properties in Barnsley

The total sample size to train the multimodal source domain is: **10,897** residential properties in Barnsley.

Numeric data

Table A.8
Statistics of numeric data of Barnsley source domain.

Variables	Mean	Std	CV
Total floor area	88.45	44.89	0.51
Number habitable rooms	2.75	2.61	0.95
Number heated rooms	4.39	1.26	0.29
Lighting description	0.67	0.37	0.55
Energy consumption (kWh)	14,509	10,149	0.69

Categorical data

Statistics of the categorical features of Barnsley source domain

Table A.9
Property type.

Property type	Count	Proportion
Bungalow	575	5.28 %
Flat	1316	12.08 %
House	8969	82.31 %
Maisonette	36	0.33 %

Table A.10
Built form.

Built form	Count	Proportion
Detached	3887	35.67 %
Enclosed End-Terrace	95	0.87 %
Enclosed Mid-Terrace	38	0.35 %
End-Terrace	1506	13.82 %
Mid-Terrace	1517	13.92 %
Semi-Detached	3566	32.73 %
unknown	287	2.63 %

Table A.11
Floor description.

Floor descriptions	Count	Proportion
Another dwelling below	861	7.90 %
Solid, insulated	2867	26.31 %
Solid, uninsulated	883	8.10 %
Suspended, insulated	1117	10.25 %
Suspended, uninsulated	766	7.03 %
To external air, insulated	23	0.21 %
To external air, uninsulated	2	0.02 %
To unheated space, insulated	33	0.30 %
To unheated space, uninsulated	18	0.17 %
Average U-Value 0–1.33	4292	39.39 %
unknown	34	0.31 %

Table A.12
Windows description.

Windows descriptions	Count	Proportion
Double glazing	7242	66.46 %
High performance glazing	3564	32.71 %
Multiple glazing	2	0.02 %
Single glazing	43	0.39 %
Triple glazing	10	0.09 %
unknown	35	0.32 %

Table A.13
Walls description.

Walls descriptions	Count	Proportion
Cavity wall, insulated	5136	47.14 %
Cavity wall, uninsulated	535	4.91 %
Cob, as built	3	0.03 %
Granite or whin, uninsulated	10	0.09 %
Sandstone or limestone, insulated	90	0.83 %
Sandstone or limestone, uninsulated	182	1.67 %
Solid brick, insulated	48	0.44 %
Solid brick, uninsulated	231	2.12 %
System built, insulated	17	0.16 %
System built, uninsulated	15	0.14 %
Timber frame, insulated	109	1.00 %
Timber frame, uninsulated	1	0.01 %
Average U-Value 0–2.1	4485	41.16 %
unknown	34	0.31 %

Table A.14
Roof description.

Roof descriptions	Count	Proportion
Another dwelling above	705	6.47 %
Flat, insulated	15	0.14 %
Flat, uninsulated	9	0.08 %
Pitched, insulated	5361	49.20 %
Pitched, uninsulated	272	2.50 %
Roof room(s), insulated	180	1.65 %
Roof room(s), uninsulated	8	0.07 %
Thatched	2	0.02 %
Average U-Value 0–2.4	4305	39.51 %
unknown	39	0.36 %

Table A.15
Mainheat descriptions.

Mainheat descriptions	Count	Proportion
Air source heat pump	78	0.72 %
Boiler	10,220	93.80 %
Community scheme	149	1.37 %
Electric heaters	126	1.16 %
Ground source heat pump	4	0.04 %
Room heaters	247	2.27 %
Warm air	5	0.05 %
Unknown	67	0.61 %

Table A.16
Construction age band.

Construction age band	Count	Proportion
1900–1920	369	3.39 %
1930–1949	215	1.97 %
1950–1966	270	2.48 %
1967–1975	223	2.05 %
1976–1982	121	1.11 %
1983–1990	210	1.93 %
1991–2002	1890	16.60 %
Before 1900	161	1.48 %
Post 2002	3253	29.85 %
Unknown	4265	39.14 %

When training for the target domain after integrating the transfer learning element, significant data loss was experienced when matching EPCs with their respective MARVEL captured building facade. The total sample size used is: **1,547** residential properties in Barnsley. The statistics are what follows:

Numeric data

Table A.17
Statistics of numeric data used for Barnsley target domain.

Variables	Mean	Std	CV
Total floor area	83.75	33.59	0.40
Number habitable rooms	4.69	1.87	0.40
Number heated rooms	4.64	1.83	0.39
Lighting description	0.52	0.36	0.69
Energy consumption (kWh)	16,309	9973	0.61

Categorical data

Statistics of the categorical features of Barnsley target domain

Table A.18
Property type.

Property type	Count	Proportion
Flat	598	38.67 %
House	949	61.33 %

Table A.19
Built form.

Built form	Count	Proportion
Detached	660	42.67 %
Enclosed End-Terrace	89	5.75 %
Enclosed Mid-Terrace	25	1.62 %
End-Terrace	278	18.00 %
Mid-Terrace	198	12.80 %
Semi-Detached	297	19.20 %

Table A.20
Floor description.

Floor descriptions	Count	Proportion
Another dwelling below	423	27.34 %
Solid, insulated	742	47.96 %
Solid, uninsulated	32	2.07 %
Suspended, insulated	258	16.67 %
Suspended, uninsulated	92	5.95 %

Table A.21
Windows description.

Windows descriptions	Count	Proportion
Double glazing	1547	100 %

Table A.22
Walls description.

Walls descriptions	Count	Proportion
Cavity wall, insulated	1279	82.66 %
Cavity wall, uninsulated	226	14.67 %
Timber frame, insulated	42	2.66 %

Table A.23
Roof description.

Roof descriptions	Count	Proportion
Another dwelling above	395	25.53 %
Flat, insulated	12	0.78 %
Pitched, insulated	1140	73.69 %

Table A.24
Mainheat descriptions.

Mainheat descriptions	Count	Proportion
Boiler	1547	100 %

Table A.25
Construction age band.

Construction age band	Count	Proportion
1930–1949	41	2.67 %
1976–1982	21	1.33 %
1991–2002	464	30.00 %
Post 2002	980	63.33 %
Unknown	41	2.67 %

A.2. Properties in Doncaster

The MARVEL capture conducted in Doncaster was only in a neighbourhood, which has limited capture. Further data loss was experienced when linking different modalities. The statistics below only represent the properties examined in this work, not the distribution for the entire housing stocks in Doncaster. The total sample size used for the Doncaster domain is: **451** residential properties in Doncaster.

Numeric data
Statistics of numeric data used for Doncaster domain

Variables	Mean	Std	CV
Total floor area	79.13	22.73	0.29
Number habitable rooms	4.25	1.01	0.24
Number heated rooms	4.25	1.01	0.24
Lighting description	1	0	0
Energy consumption (kWh)	8228	4431	0.54

Categorical data
Statistics of the categorical features of Doncaster target domain

Table A.26
Property type.

Property type	Count	Proportion
Bungalow	263	58.33 %
House	188	41.67 %

Table A.27
Built form.

Built form	Count	Proportion
Detached	19	4.17 %
End-Terrace	75	16.67 %
Mid-Terrace	56	12.50 %
Semi-Detached	301	66.67 %

Table A.28
Floor description.

Floor descriptions	Count	Proportion
Solid, uninsulated	319	71.7 %
Average U-Value 0–1.33	132	29.3 %

Table A.29
Windows description.

Windows descriptions	Count	Proportion
Double glazing	38	8.33 %
High performance glazing	413	91.67 %

Table A.30
Walls description.

Walls descriptions	Count	Proportion
Cavity wall, insulated	198	43.90 %
Solid brick, uninsulated	205	45.5 %
Average U-Value 0–2.1	48	10.6 %

Table A.31
Roof description.

Roof descriptions	Count	Proportion
Pitched, insulated	192	42.6 %
Average U-Value 0–2.4	259	57.4 %

Table A.32
Mainheat descriptions.

Mainheat descriptions	Count	Proportion
Boiler	451	100 %

Table A.33
Construction age band.

Construction age band	Count	Proportion
1976–1982	207	45.90 %
Before 1900	244	54.10 %

A.3. Properties in Merthyr Tydfil

Similar to other target regions, data loss due to matching modalities was experienced in Merthyr Tydfil and resulted in a smaller sample size.

Total sample size: **1,345** residential properties in Merthyr Tydfil.

Numeric data

Table A.34
Statistics of numeric data used for Merthyr Tydfil target domain.

Variables	Mean	Std	CV
Total floor area	91.64	11.71	0.13
Number habitable rooms	4.32	0.63	0.15
Number heated rooms	4.232	0.63	0.15
Lighting description	0.60	0.20	0.33
Energy consumption (kWh)	14,203	5304	0.37

Categorical data

Statistics of the categorical features of Merthyr Tydfil target domain.

Table A.35
Property type.

Property type	Count	Proportion
Flat	40	2.94 %
House	1305	97.06 %

Table A.36
Built form.

Built form	Count	Proportion
End-Terrace	237	17.65 %
Mid-Terrace	316	23.53 %
Semi-Detached	792	58.82 %

Table A.37
Floor description.

Floor descriptions	Count	Proportion
Solid, insulated	521	38.71 %
Solid, uninsulated	695	51.65 %
Suspended, uninsulated	129	9.64 %

Table A.38
Windows description.

Windows descriptions	Count	Proportion
Double glazing	1345	100 %

Table A.39
Walls description.

Walls descriptions	Count	Proportion
Cavity wall, insulated	923	68.52 %
Cavity wall, uninsulated	80	5.95 %
System built, insulated	342	25.43 %

Table A.40
Roof description.

Roof descriptions	Count	Proportion
Another dwelling above	292	21.71 %
Pitched, insulated	761	56.58 %
Pitched, uninsulated	292	21.71 %

Table A.41
Mainheat descriptions.

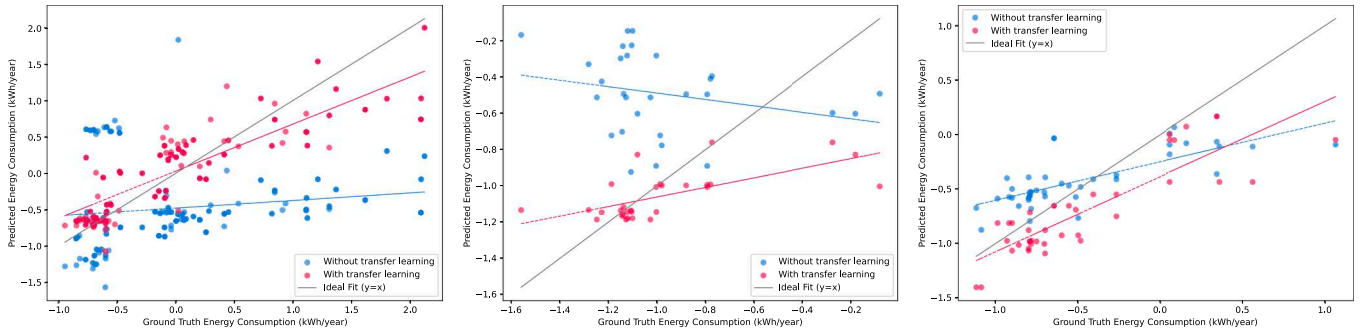
Mainheat descriptions	Count	Proportion
Boiler	1345	100 %

Table A.42
Construction age band.

Construction age band	Count	Proportion
1930–1949	158	11.76 %
1950–1966	119	8.82 %
1991–2002	198	14.71 %
Unknown	870	64.71 %

Appendix B. Comparisons between the prediction results with and without transfer learning component

Further examination of the prediction results is presented. The ground truth energy consumption data extracted from EPC were plotted against the predicted value using the proposed multimodal transfer learning network. Linear regressions were used to illustrate the



(a) Barnsley (b) Doncaster (c) Merthyr Tydfil

Fig. B.1. Scatter plots of the ground truth annual energy consumption (EPC) against the predicted annual energy consumption.

relationship. As presented in Fig. B.1, the resulted plots further revealed that the proposed incorporation of transfer learning has significantly improved the prediction performance. However, the trendlines suggest the proposed models tend to underestimate energy consumption relative to EPC ground-truth values.

References

[1] BEIS, UK Sets Ambitious New Climate Target Ahead of UN Summit, Technical Report, BEIS, 2020. https://www.gov.uk/government/news/uk-sets-ambitious-new-climate-target-ahead-of-un-summit.

[2] DESNZ, BEIS, Accredited official statistics: energy consumption in the UK 2023, 2024, https://www.gov.uk/government/statistics/energy-consumption-in-the-uk-2023, visited on 2024-09-03.

[3] GOV.UK, Apply for help to improve a home with no gas boiler (Home Upgrade Grant), 2024, https://www.gov.uk/apply-home-upgrade-grant, visited on 2024-09-03.

[4] BEIS, The Heat and Buildings Strategy, 2021, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1044598/6.7408_BEIS_Clean_Heat_Heat__Buildings_Strategy_Stage_2_v5_WEB.pdf.

[5] J. Alabid, A. Bennadji, M. Seddiki, A review on the energy retrofit policies and improvements of the UK existing buildings, challenges and benefits, Renew. Sustain. Energy Rev. 159 (2022) 112161.

[6] D.V. Carvalho, E.M. Pereira, J.S. Cardoso, Machine learning interpretability: a survey on methods and metrics, Electronics 8 (2019). https://doi.org/10.3390/electronics8080832

[7] C. Molnar, Interpretable Machine Learning, Lulu. com, 2020.

[8] P. Linardatos, V. Papastefanopoulos, S. Kotsiantis, Explainable AI: a review of machine learning interpretability methods, Entropy 23 (2021) 1–45. https://doi.org/10.3390/e23010018

[9] P.J.G. Lisboa, S. Saralajew, A. Vellido, R. Fernández-Domenech, T. Villmann, The coming of age of interpretable and explainable machine learning models, Neurocomputing 535 (2023) 25–39. https://doi.org/10.1016/j.neucom.2023.02.040

[10] W. Li, Y. Zhou, K. Cetin, J. Eom, Y. Wang, G. Chen, X. Zhang, Modeling urban building energy use: a review of modeling approaches and procedures, Energy 141 (2017) 2445–2457.

[11] M. Bourdeau, X.q. Zhai, E. Nefzaoui, X. Guo, P. Chatellier, Modeling and forecasting building energy consumption: a review of data-driven techniques, Sustain. Cities Soc. 48 (2019). https://doi.org/10.1016/j.scs.2019.101533

[12] H.-X. Zhao, F. Magoulès, A review on the prediction of building energy consumption, Renew. Sustain. Energy Rev. 16 (6) (2012) 3586–3592.

[13] R. Olu-Ajayi, H. Alaka, I. Sulaimon, F. Sunmola, S. Ajayi, Building energy consumption prediction for residential buildings using deep learning and other machine learning techniques, J. Build. Eng. 45 (2022). Visited on 2024-01-13. https://doi.org/10.1016/j.jobe.2021.103406

[14] U. Ali, M.H. Shamsi, F. Alshehri, E. Mangina, J. O'Donnell, Application of intelligent algorithms for residential building energy performance rating prediction, Building Simulation Conference Proceedings 5 (2019) 3177–3184. https://doi.org/10.26868/25222708.2019.210232

[15] A.A.A. Gassar, G.Y. Yun, S. Kim, Data-driven approach to prediction of residential energy consumption at urban scales in London, Energy 187 (2019). https://doi.org/10.1016/j.energy.2019.115973

[16] M. Despotovic, D. Koch, S. Leiber, M. Döller, M. Sakeena, M. Zeppelzauer, Prediction and analysis of heating energy demand for detached houses by computer vision, Energy Build. 193 (2019) 29–35. https://doi.org/10.1016/j.enbuild.2019.03.036

[17] U. Ali, M.H. Shamsi, M. Bohacek, C. Hoare, K. Purcell, E. Mangina, J. O'Donnell, A data-driven approach to optimize urban scale energy retrofit decisions for residential buildings, Appl. Energy 267 (2020) 114861. https://doi.org/10.1016/j.apenergy.2020.114861

[18] M. Sun, C. Han, Q. Nie, J. Xu, F. Zhang, Q. Zhao, Understanding building energy efficiency with administrative and emerging urban big data by deep learning in Glasgow, Energy Build. 273 (2022). Visited on 2024-01-13. https://doi.org/10.1016/j.enbuild.2022.112331

[19] C.E. Kontokosta, C. Tull, A data-driven predictive model of city-scale energy use in buildings, Appl. Energy 197 (2017) 303–317. Visited on 2024-01-13. https://doi.org/10.1016/j.apenergy.2017.04.005

[20] S. Wenninger, C. Wiethe, Benchmarking energy quantification methods to predict heating energy performance of residential buildings in Germany, Bus. Inf. Syst. Eng. 63 (2021) 223–242. https://link.springer.com/article/10.1007/s12599-021-00691-2, visited on 2024-01-13.

[21] ONS, Age of the Property is the Biggest Single Factor in Energy Efficiency of Homes, Technical Report, 2022. https://www.ons.gov.uk/peoplepopulationandcommunity/housing/articles/ageoftheproperty isthebiggest singlefactorinenergyefficiencyofhomes/2021-11-01, visited on 2024-01-13.

[22] D. Jenkins, S. Simpson, A. Peacock, Investigating the Consistency and Quality of EPC Ratings and Assessments, Energy 138 (2017) 480–489. visited on 2024-01-13. https://doi.org/10.1016/j.energy.2017.07.105

[23] D.o. Energy, C. Change, Green Deal Assessment Mystery Shopping Research, Technical Report, Department of Energy and Climate Change, 2014. https://www.gov.uk/government/publications/green-deal-assessment-mystery-shopping-research, visited on 2024-01-13.

[24] J. Summaira, X. Li, A.M. Shoib, S. Li, J. Abdul, Recent advances and trends in multimodal deep learning: a review, arXiv preprint (2021). http://arxiv.org/abs/2105.11087.

[25] K. Bayoudh, R. Knani, F. Hamdaoui, A. Mtibaa, A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets, Visual Comput. 38 (2022) 2939–2970. https://doi.org/10.1007/s00371-021-02166-7

[26] Y. Sheng, W.O.C. Ward, H. Arbabi, M. Álvarez, M. Mayfield, Deep multimodal learning for residential building energy prediction, in: IOP Conference Series: Earth and Environmental Science, 1078, IOP Publishing, 2022, p. 012038.

[27] F. Biljecki, K. Ito, Street view imagery in urban analytics and GIS: a review, Landsc. Urban Plan. 215 (2021) 104217.

[28] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017).

[29] A. Hooshmand, R. Sharma, Energy predictive models with limited data using transfer learning, in: e-Energy 2019 - Proceedings of the 10th ACM International Conference on Future Energy Systems, Association for Computing Machinery, Inc, 2019, pp. 12–16. https://doi.org/10.1145/3307772.3328284

[30] Y. Gao, Y. Ruan, C. Fang, S. Yin, Deep learning and transfer learning models of energy consumption forecasting for a building with poor information data, Energy Build. 223 (2020) 110156. https://doi.org/10.1016/j.enbuild.2020.110156

[31] A. Farahani, S. Voghoei, K. Rasheed, H.R. Arabnia, A brief review of domain adaptation, Advances in Data Science and Information Engineering: Proceedings from ICDATA 2020 and IKE 2020 (2021) 877–894.

[32] Z. Zhao, L. Alzubaidi, J. Zhang, Y. Duan, Y. Gu, A comparison review of transfer learning and self-supervised learning: definitions, applications, advantages and limitations, Expert Syst. Appl. 242 (2024) 122807.

[33] Q. Zhang, J. Niu, Z. Tian, L. Bao, J. Luo, M. Wang, Y. Cao, A study on source domain selection for transfer learning-based cross-building cooling load prediction, Energy Build. 324 (2024) 114856.

[34] X. Fang, G. Gong, G. Li, L. Chun, W. Li, P. Peng, A hybrid deep transfer learning strategy for short term cross-building energy prediction, Energy 215 (2021). https://doi.org/10.1016/j.energy.2020.119208

[35] H. Jin, F. Chollet, Q. Song, X. Hu, AutoKeras: an AutoML library for deep learning, J. Mach. Learn. Res. 24 (6) (2023) 1–6.

[36] G. Meyers, C. Zhu, M. Mayfield, D.D. Tingley, J. Willmott, D. Coca, Designing a vehicle mounted high resolution multi-spectral 3d scanner: Concept design, in: Proceedings of the 2nd Workshop on Data Acquisition to Analysis, 2019, pp. 16–21.

[37] B. Lartigue, L. Biewesch, F. Marion, E. Cochran, F. Thellier, Energy performance certificates in the USA and in France–A case study of multifamily housing, Energy Effic. 15 (2022). https://doi.org/10.1007/s12053-022-10036-x

[38] F. Biljecki, K. Ito, Street view imagery in urban analytics and GIS: a review, 2021, https://doi.org/10.1016/j.landurbplan.2021.104217

[39] M. Zeppelzauer, M. Despotovic, M. Sakeena, D. Koch, M. Döller, Automatic prediction of building age from photographs, in: Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, 3206060, ACM, 2018, pp. 126–134. https://doi.org/10.1145/3206025.3206060

[40] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-Decem, 2016, pp. 779–788. https://pjreddie.com/darknet/yolo/. https://doi.org/10.1109/CVPR.2016.91

[41] X. He, K. Zhao, X. Chu, AutoML: a survey of the state-of-the-art, Knowl. Based Syst. 212 (2021) 106622.

[42] F. Hutter, L. Kotthoff, J. Vanschoren, Automated Machine Learning: Methods, Systems, Challenges, Springer Nature, 2019.

[43] B. Nikparvar, J.-C. Thill, Machine learning of spatial data, ISPRS Int. J. Geoinf. 10 (9) (2021) 600.

[44] J. Crawley, P. Biddulph, P.J. Northrop, J. Wingfield, T. Oreszczyn, C. Elwell, Quantifying the measurement error on England and Wales EPC ratings, Energies 12 (2019). https://doi.org/10.3390/en12183523

[45] A. Hardy, D. Glew, An analysis of errors in the energy performance certificate database, Energy Policy 129 (2019) 1168–1178. https://doi.org/10.1016/j.enpol.2019.03.022

[46] P. Tang, D. Huber, B. Akinci, R. Lipman, A. Lytle, Automatic reconstruction of as-built building information models from laser-scanned point clouds: a review of related techniques, Autom. Constr. 19 (7) (2010) 829–843.

[47] Y. Xu, U. Stilla, Toward building and civil infrastructure reconstruction from point clouds: a review on data and key techniques, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14 (2021) 2857–2885.

[48] U.B.D. Centre, Glasgow 3D city models derived from airborne LiDAR point clouds licensed data, 2024, https://doi.org/10.20394/VWYL2ON6

[49] S. Salmani Pour Avval, N.D. Eskue, R.M. Groves, V. Yaghoubi, Systematic review on neural architecture search, Artif. Intell. Rev. 58 (3) (2025) 73.