Energy & Buildings 367 (2026) 117740

Contents lists available at ScienceDirect

Energy & Buildings

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

Publicly accessible data and computer vision for retrofit planning: Evidence from London social housing

Robin Roussel [a,*], Luke Gooding [b]

[a] School of Architecture, Royal College of Art, Darwin Building, Kensington Gore, London, SW7 2EU, United Kingdom
[b] Stockholm Environment Institute, University of York, Environment Building, Wentworth Way, York, YO10 5NG, United Kingdom

ARTICLE INFO

Keywords:
Energy retrofit
Social housing
Computer vision
Building data

ABSTRACT

Buildings account for approximately 34% of global CO_2 emissions, placing the built environment at the centre of climate action. Ageing social housing represents a particularly under-studied challenge: in the United Kingdom, around 78% of the building stock predates 1980, and nearly 60% of buildings carry poor energy performance ratings. Systematic retrofit planning is further hampered by building data that is incomplete and fragmented across disparate sources. This paper presents a pilot study investigating data-driven approaches to support social housing retrofit planning across several London boroughs. Rather than proposing a deployable pipeline or a novel neural-network architecture, its purpose is to test what can be inferred from currently accessible property and image data, and to identify the technical and organisational difficulties that arise when doing so. Using publicly accessible property datasets, computer vision training data, and street-level imagery, convolutional neural networks were fine-tuned to predict two retrofit-relevant building characteristics: wall construction type and architectural style. These two tasks were selected as diagnostic cases: wall construction is directly relevant to fabric-first retrofit decisions, while architectural style provides a visually observable proxy for construction age and illustrates the opportunities and risks of using existing non-retrofit image datasets. In-distribution (ID) accuracy reached up to 94%; however, performance declined substantially when models were evaluated against out-of-distribution (OOD) property samples provided by local councils. A structured analysis of the ID-OOD image domain shift revealed a gap in terms of pixel statistics, frequency profile, and semantic content that is consistent with the performance drop for each task. This highlights the generalisation challenges inherent in applying models trained on publicly accessible data to real-world contexts, and underscores the need for broader, purpose-built labelled datasets. A stakeholder workshop with representatives from local government, retrofit organisations, and industry consultancies complemented the technical findings. Discussions surfaced practical barriers to data use in the retrofit sector, including inconsistent data formats, limited inter-organisational data sharing, and a lack of coordination across the supply chain. The findings suggest that publicly accessible housing data and machine learning hold genuine promise for large-scale urban retrofit characterisation, but that realising this potential will require greater data standardisation, coordinated investment in labelled datasets, and closer collaboration across the sector.

1. Introduction

Decarbonising the existing building stock has now been recognised as an important requirement for meeting the targets for climate mitigation. Globally, the buildings and construction sector has been estimated to contribute around 34% of the total energy-related CO_2 emissions. This underlines the importance of the built environment in the context of climate change mitigation [1]. On the other hand, the turnover rates of buildings are low. This means that the buildings that are expected to be around in 2050 have already been built. Hence, the focus has to be more on the decarbonisation of existing buildings. This is particularly the case in the United Kingdom, where the existing housing stock is considered old and heterogeneous. For instance, a considerable number of the existing housing stock was built before the implementation of modern building regulations. It is estimated that around 78% of the existing housing stock was built before the year 1980 [2]. Moreover, a considerable number of the existing housing stock has poor energy efficiency ratings, with many homes scoring less than band C [3].

Within this context, social housing provision is a key area. Social housing providers manage large quantities of housing stock and

* Corresponding author.
E-mail addresses: robin.roussel@ucl.ac.uk (R. Roussel), luke.gooding@york.ac.uk (L. Gooding).

https://doi.org/10.1016/j.enbuild.2026.117740
Received 30 March 2026; Received in revised form 13 May 2026; Accepted 30 May 2026
Available online 9 June 2026

may be under a policy obligation with respect to decarbonisation and fuel poverty reduction. However, retrofit delivery in the social housing sector is also influenced by a range of constraints [4,5]. Therefore, retrofit delivery needs to balance feasibility, cost, complexity, and governance.

1.1. Data challenges in retrofit planning

Thus, effective retrofit planning relies on the availability of reliable and comprehensive data on the characteristics of the buildings. However, the data on the characteristics of the buildings is fragmented, incomplete, and often not easily integrated. Information relevant to retrofit planning and delivery, such as the construction of the external walls of the buildings, or their age, geometry, and occupancy, is usually found in different datasets, including Energy Performance Certificates (EPCs), local government data, survey data, and geospatial data [6,7]. EPCs are one of the most accessible sources of property-level data in the UK, as well as one of the most commonly used in the analysis of the building stock. However, there is a general consensus about the limitations of the EPC data, which is not comprehensive and collected only at the time of a trigger event like the sale or rental of a property [8,9]. In addition, the data is based on standard modelling assumptions that may not reflect the real-world situation [10]. While local government data may be occasionally available, it is not in an easily accessible and standardised form. Data sharing between organisations is frequently constrained by technical, institutional, and legal barriers, resulting in what has been described as a "patchwork" of retrofit-relevant information [6]. As a result, local authorities and housing providers often face significant challenges in prioritising buildings, defining retrofit archetypes, and designing cost-effective interventions.

1.2. Addressing the data challenges

Several broad approaches have been proposed to respond to these problems. First, building-stock models [11] and retrofit archetypes [12] can reduce complexity by grouping properties with similar physical characteristics. Second, urban data platforms and digital twins can support data integration across administrative, geospatial, and physical datasets [13]. Third, machine learning methods can help infer missing or inconsistent attributes from visual or spatial data [14]. This study focuses on the third approach while recognising that it is only useful when embedded within the first two: automated inference must feed into archetyping and data-governance workflows rather than operate as a stand-alone substitute for professional assessment.

Among the digital and data-driven approaches for building stock analysis, computer vision is seen as an emerging and promising technology. This is due to the ability to leverage recent advancements in deep learning to extract useful information from visual data sources such as satellite imagery, aerial photography, and street-level imagery [14–16]. Computer vision has been effectively used for building typology classification, façade element recognition, and estimation of building physical characteristics [14,17]. Street-level imagery is particularly relevant in this case since it is able to capture information about the façade of the buildings, which is not easily attainable through other data sources. Past studies have proven that it is possible to use the imagery to make inferences about the buildings, their age, materiality, and condition [16]. This implies that computer vision has the potential to fill in the gaps in the data related to buildings, particularly in cases where the data is not easily accessible. However, the potential of computer vision in the field of retrofit planning is not adequately explored. Most of the existing research is based on the technical capability of computer vision in a controlled environment, rather the usability of the results of machine learning in the planning process [14]. In addition, the issue of generalisation and data bias in the results of computer vision for retrofit applications is not sufficiently studied.

The present study therefore uses a deliberately simple and transparent computer vision pipeline. It does not claim to introduce a new neural-network architecture. Instead, this work uses a standard ResNet-50 backbone [18] as a controlled test case to examine whether publicly accessible labels and imagery can support retrofit-relevant classification, and where such an approach breaks down when moved from benchmark-style data to municipal retrofit samples.

1.3. Research gap

The literature on building retrofit and energy performance has developed robust approaches to archetype modelling, techno-economic analysis and policy evaluation [19,20]. At the same time, research on digital built environment systems has advanced rapidly, particularly in the areas of machine learning and urban data analytics [16]. However, relatively few studies have combined these perspectives in a way that addresses practical retrofit planning challenges.

In particular, there is a lack of research that integrates:

(i) automated detection of retrofit-relevant building characteristics,
(ii) publicly available property-level and visual datasets, and
(iii) validation against real-world municipal retrofit data and workflows.

This gap is significant because the effectiveness of smart technologies in the built environment depends not only on model accuracy, but also on their ability to operate under the data constraints and institutional conditions of retrofit delivery. As noted in recent studies, the generalisation of machine learning models across different urban contexts remains a key challenge, particularly where training datasets are limited or unrepresentative [14,16].

The study also responds to a more specific gap concerning the relationship between public data availability and retrofit implementation. A dataset may be technically accessible while still being unsuitable for direct deployment because it is incomplete, inconsistently labelled, governed by restrictive licences, or misaligned with the geography and building types targeted by local authorities. This distinction between accessibility and practical usability is central to the contribution of the paper.

1.4. Aim and contributions

This study addresses this gap by investigating whether publicly available housing datasets, combined with computer vision techniques, can support retrofit planning at scale in the context of social housing. The paper makes four main contributions. First, it develops a data integration pipeline linking EPC data and street-level imagery at the property level. Second, it applies computer vision models to estimate wall type and construction age proxies relevant to retrofit archetyping. Third, it evaluates model performance against local authority retrofit assessment data, providing insight into real-world applicability and generalisation challenges. Finally, it complements the technical analysis with stakeholder insights on data governance and practical barriers to data use in retrofit planning.

The contribution is empirical and diagnostic rather than architectural. The study does not propose a deployment-ready pipeline for retrofit stakeholders, nor does it attempt to classify the full set of attributes needed for retrofit design. Instead, it uses two bounded classification tasks to identify what can be learned from accessible datasets and to expose the difficulties that would need to be addressed before such methods could support operational retrofit planning.

2. Literature review

2.1. Domestic retrofit and energy performance

Retrofitting of existing buildings is one area that is currently receiving considerable attention, both in research and policy, because of its critical role towards achieving decarbonisation and energy demand reduction. Residential buildings have a high share of operational energy use, and their retrofit is considered one of the most cost-effective ways of achieving emission reductions in the short to medium term [1,21]. However, achieving this goal is particularly difficult in situations where there is a high prevalence of ageing buildings as well as varied construction types and building ownership structures. Such challenges abound in the UK social housing sector. Social housing providers are required to achieve ever more demanding energy performance requirements while also ensuring the affordability and well-being of the residents. Moreover, retrofit interventions are required to be implemented within tight financial constraints and sometimes in an environment of policy uncertainty [4,5]. Empirical research has demonstrated that the adoption of retrofit interventions is influenced by a variety of technical, institutional, and behavioural factors [4,22].

One important aspect in retrofit research is the so-called “performance gap.” Various studies have shown that there is a significant discrepancy between modelled and actual energy savings due to various reasons such as occupant behaviour and construction quality [10,23]. The performance gap is important in policy and practical contexts because it adds a degree of complexity in cost-benefit analysis and evaluation of retrofit schemes. It also points to the limitations of using standardised assessment tools – such as the Reduced data Standard Assessment Procedure (RdSAP) [24] – in planning and evaluation. To cope with the complexity of the housing stock, researchers as well as experts often apply the idea of “retrofit archetypes”. Archetypes are used as a tool for grouping buildings with comparable characteristics, facilitating the assessment of the interventions on a large scale, as opposed to the level of the buildings themselves [12,25]. The idea of archetypes is the basis of many building stock models as well as tools used in policy development. Nevertheless, the success of the archetyping process largely depends on the availability as well as the accuracy of the data on the characteristics of the buildings, including the age, type of buildings, as well as the condition of the building fabric [19]. Overall, the literature implies that while retrofit archetypes and modelling approaches are clearly important tools in the process of scaling up, their utility is limited in certain ways by data availability and the inherent variability of real-world buildings. Addressing these limitations, however, will require better tools for characterising buildings and integrating diverse data sets.

Recent work has emphasised that the usefulness of stock-level modelling depends on the ability to link building data with wider urban systems and heterogeneous urban datasets. For example, transportation-informed occupancy estimation has been used to improve urban-scale building occupancy and energy-use estimates by incorporating mobility-derived information into building energy modelling [26]. This is relevant to retrofit planning because it demonstrates the value of fusing building attributes with broader geospatial and infrastructural information when direct observations are incomplete. Accordingly, the present work is positioned within a broader shift from single-source building datasets toward multi-source urban inference.

2.2. Data-driven approaches for the built environment

The limitation of conventional data sources has created increasing interest in data-driven techniques to develop building stock analysis. Urban building energy modelling is one such area of research, where building energy models have incorporated geospatial data, statistical, and machine learning techniques to study building stocks at the urban and regional levels [27]. Digital twins and urban data platforms are also being considered as important tools in this domain. Digital twins are defined as an effort to create virtual copies of physical assets and systems by integrating data from different sources for monitoring and decision-making purposes [28]. In the context of buildings, the concept of digital twins could integrate geometric data, sensor data, and other information for the purpose of providing insights. The application of this concept is still a challenge due to the problems associated with the availability and cost of data. Urban data platforms are, in part, a response to these issues, as they offer shared infrastructures for data storage, integration, and analysis, which can help support the aggregation of data from buildings across jurisdictions, allowing for more coordinated approaches to planning and policy [6]. However, the success of these data platforms is reliant on the quality of the data used, as well as the governance structures in support of data sharing. In this context, computer vision has attracted more attention as a technique for extracting building information from images. This is because of the advancement of deep learning, which allows for the classification of building types, detection of building façade components, and estimation of building physical attributes from images [14]. Street view imagery (SVI), for example, has been employed for the estimation of building age, building condition, and socio-economic status [16].

Applications of computer vision for architecture and planning have shown the possibility for automating certain aspects of building analysis that otherwise need manual analysis [14]. For instance, convolutional neural networks have been used on SVI to classify building typologies [29] and architectural styles [17,30]. Other applications have also focused on building façade segmentation for identifying building features such as windows and doors [31]. Despite these developments, however, some challenges need to be overcome. These include the sensitivity of the models to differences in image quality and context, as well as the need for generalisability to be ensured [15]. Moreover, most studies have focused on the technical classification aspects rather than the ways in which the output can be used in decision-making. Hence, there is a need for studies that link computer vision to retrofit planning and assess their utility.

A further strand of relevant research concerns the use of multi-view imagery and graph-based modelling to improve building characterisation. Such approaches are relevant to retrofit applications because building attributes such as wall construction, roof type, and glazing ratio may be only partially visible from a single viewpoint [32,33]. For instance, aerial and street view images have been combined to directly predict residential electricity consumption [34]. SVI has also been fused with remote sensing data (including imagery and land surface temperature) to estimate building energy efficiency [35] and identify hard-to-decarbonise homes [36]. These developments suggest that future retrofit-characterisation systems may even need to combine visual inference with geospatial and tabular data sources rather than relying on image content alone. The present pilot adopts a simpler single-property classification setup, but its findings on domain shift and data integration help illustrate why more integrated approaches may ultimately be required.

2.3. Data governance and participation in retrofit

However, it has also been recognised that technical solutions alone are not sufficient to overcome the challenges associated with retrofit planning. This is particularly due to the fact that the governance of data also plays a critical role in how information is generated, shared, and utilised. For instance, in the built environment, there are often multiple organisations involved in managing data, including government agencies, housing associations, consultants, and private companies. This has been identified as a challenge to integration and how effectively stakeholders can come to a shared understanding of building stock characteristics [6]. Publicly available data such as EPCs have also been identified as a partial solution to this problem, although their limitations have also been well documented. For instance, there are concerns regarding their completeness, quality, and lack of standardisation between records [8]. Moreover, there are also concerns regarding how data is often collected for administrative purposes rather than strategic planning, which can lead to a lack of integration between datasets when repurposed for retrofit analysis. In the case of local authorities, the data challenges relate to decision-making. In retrofit planning, there is a need for interdepartmental coordination and the involvement of external actors. Research has indicated that local authorities may not have the capacity to fully harness the potential of the data at their disposal. This may result in the hiring of external consultants to manage the data [5]. Another factor in retrofit governance is participation. The retrofit programme has a direct impact on the people. However, the people may not be involved in the decision-making process. In recent times, there has been a need to involve the people in the decision-making process not only as recipients of the retrofit programme but also as producers of knowledge [4]. This is because the people may have an in-depth understanding of the retrofit programme. In this context, there has been an increasing interest in the development of participatory data infrastructures that integrate technical data sources with local knowledge and user engagement. These approaches acknowledge the social and institutional construction of data and the importance of technical and social forms of evidence for effective retrofit planning [22]. Nevertheless, the implementation of participatory data systems also poses other challenges with regard to data governance and privacy issues.

In conclusion, the literature review has demonstrated that effective retrofit planning requires not just more and better data and more effective data analysis and interpretation, but also more effective data governance and more participatory approaches. Based on the gaps identified in the literature, this study addresses the following research questions:

RQ1. To what extent can publicly accessible data sources support the construction of supervised training datasets for retrofit-relevant building characteristics?
RQ2. What is the performance and cross-domain generalisation of computer vision models for predicting retrofit-relevant building characteristics when trained on (a) relevant labels derived from public registries, or (b) datasets not purpose-built for retrofit analysis?
RQ3. What challenges arise when applying these models in real-world municipal retrofit planning contexts?

3. Methodology

This study employs a mixed-methods design comprising two sequential phases: a development pipeline to identify opportunities and limitations in applying machine learning to building characterisation, followed by a two-step evaluation stage combining a quantitative assessment and a stakeholder workshop to contextualise the findings within broader retrofit efforts and explore pathways to implementation.

The connection between the two evaluation steps is intentional. The first step generated quantitative evidence about model performance, generalisation failure, and data-integration constraints. The workshop stage then used these findings as prompts for structured discussion with practitioners, focusing on whether such outputs could be trusted, how they might be incorporated into workflows, and what organisational barriers would prevent deployment. The mixed-methods design therefore links model evaluation to implementation feasibility: the quantitative results identify what the models can and cannot infer, while the qualitative stage identifies what would be required for such inference to become useful in practice.

Fig. 1 presents the overall architecture of the study. The development pipeline begins with a machine learning workflow divided into task definition (Section 3.1), training data collection (Section 3.2), model selection, and training (Section 3.3). A full prediction pipeline infers the characteristics of council properties (Section 3.4). The evaluation stage includes area selection and data collection (Section 3.5), followed by quantitative assessment (Section 3.6), which feeds into workshop discussions (Section 3.7).

3.1. Machine learning tasks and building characteristics

Two machine learning tasks were selected, each linked to building features relevant to retrofit archetypes. The tasks were chosen to highlight complementary aspects and challenges associated with publicly available data and computer vision (CV) methods:

- Wall construction classification, which is directly relevant to fabric-first retrofit works.
- Architectural style classification, serving as a proxy for construction age.

Wall construction is essential for fabric-first retrofit works and forms part of the RdSAP inputs [24], from which it carries over to EPCs. Although the wall construction type is already present in publicly available EPC data, there are two motivations for developing a predictive computer vision model of this characteristic.

First, wall construction provides a well-suited test case for a CV model applied to a characteristic that is visually non-trivial yet can nonetheless be inferred by expert assessors. Visible brickwork is a commonly used diagnostic cue (see Fig. 2): English and Flemish bonds typically indicate a solid wall, whereas a stretcher bond suggests a cavity wall. In practice, however, the diversity of the building stock precludes the application of any simple rule. In rare cases, English or Flemish-bond aesthetics may have been achieved on a cavity wall extension using snapped headers – bricks cut in half – to match the appearance of the adjacent structure. Furthermore, brickwork is not always externally visible; neighbouring properties may expose bare brick, and broader inferences about construction age derived from architectural style (described below) can provide supplementary support. Second, the presence of wall construction data in standardised, publicly available sources makes it straightforward to generate training labels and conduct a rigorous evaluation – both critical requirements for a short-term pilot study. Despite well-documented quality concerns with EPC data more generally, insights from the stakeholder workshops suggest that social housing EPCs tend to be of higher quality, as local councils typically commission professional surveyors to conduct assessments. In summary, while wall construction inference from street-level imagery has limited stand-alone utility for social housing applications, it is of considerable interest as a pilot task: it enables a focused investigation of image quality, model performance, and training challenges, while minimising the confounding effects of label noise.

Architectural style does not directly inform retrofit decisions in the general case – provided the property is neither listed nor situated within a conservation area, the latter being a designation available through publicly accessible data. The rationale for including it in this study is therefore distinct from that of wall construction. Rather than adapting CV methods to a feature grounded in established retrofit standards, this task takes the inverse approach: it examines a characteristic that has received attention in the CV literature and for which datasets and pre-trained models have previously been developed [17,30], and evaluates its potential as a proxy for construction age – a strong predictor of building characteristics and energy performance [27]. Construction age itself is not a visually discernible feature; architectural style, by contrast, is inherently visual. It should be noted that SAP construction age bands were not defined to correspond to stylistic periods but rather to amendments in Building Regulations [24]. While architectural styles are not perfectly delineated – hybrid forms are common – they represent fundamentally visual categories, which accounts for their prominence in the CV literature. Quantifying the precise correlation between architectural style and construction age lies beyond the scope of this work; the primary goal is to identify challenges and opportunities in using style as a proxy. Crucially, this task enables an examination of model generalisation issues that arise from training on publicly available data, independently of image quality considerations. Although the two tasks draw on different labelling sources – EPCs for wall construction and a non-energy third-party dataset for architectural style – they are not entirely uncorrelated, as solid-wall construction is more prevalent in older architectural styles. Accordingly, a sub-question of RQ2 investigates whether supplying the output of the frozen, fine-tuned architectural style classifier as an

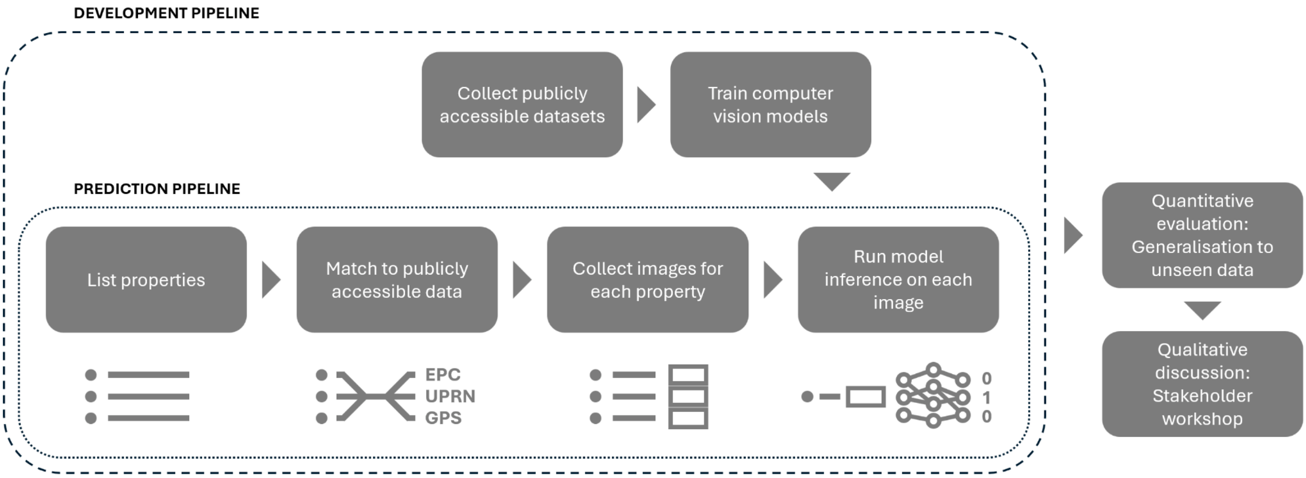

Fig. 1. Overall study architecture linking the development pipeline (machine learning workflow and prediction pipeline) with its subsequent evaluation (generalisation ability, stakeholder workshop interpretation).

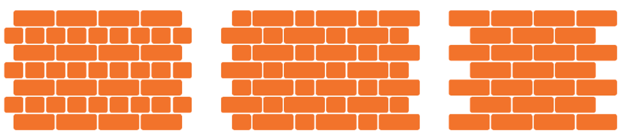

Fig. 2. Examples of typical brick bonds found in the UK: English (left), Flemish (middle), and stretcher (right).

Fig. 3. Example images used for wall construction classification, from Roussel et al. [32]. ©2019 Google.

additional input to the wall construction classifier – during both training and inference – affects the latter's performance.

3.2. Collection of publicly accessible training data

For wall construction, a pre-existing dataset of Google Street View (GSV) images of London houses from Roussel et al. [32] (see Fig. 3) was combined with wall construction information extracted from the corresponding EPCs. Although the image dataset is not technically in the public domain – owing to GSV licensing restrictions – it is accessible upon request and was originally compiled from publicly accessible sources. Its use in this study was motivated by the practical constraints of the pilot timeline. As a multi-view dataset, it yielded 738 images of 289 individual London properties, spanning three wall construction classes: solid brick (n = 542), cavity wall (n = 167), and timber frame (n = 29).

For architectural style, imagery and labels from Lindenthal and Johnson [30] were used. This dataset comprises GSV images of approximately 25,000 residential properties in Cambridge (see Fig. 4); in practice, however, only 6345 labelled images were available online. The dataset encompasses eight style classes, spanning Georgian through Contemporary, with one revival class (Revival Victorian).

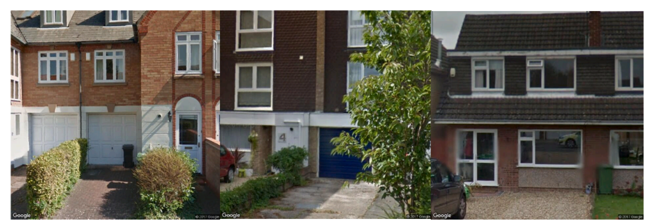

Fig. 4. Example images used for architectural style classification, from Lindenthal and Johnson [30]. ©2017 Google.

The distinction between "publicly accessible" and "openly reusable" is important. EPC records are publicly accessible and can be downloaded or queried, although they remain subject to data-quality limitations. GSV imagery is publicly accessible through platform interfaces and APIs but is not open data in the same sense: licensing terms restrict redistribution and may change over time. Consequently, this study does not claim to release a fully open image dataset. Instead, it evaluates whether data sources that are practically accessible to researchers and local authorities can support retrofit-relevant inference.

3.3. Computer vision models

For both image classification tasks – wall construction and architectural style – a well-established convolutional neural network (CNN) architecture, ResNet-50 [18], was selected. The model was initialised with weights pre-trained on ImageNet-1k [37] and subsequently fine-tuned for each task. To evaluate the scenario in which the wall construction classifier incorporates the architectural style classifier's output as an auxiliary input, a module was appended to the end of the ResNet backbone. This module projects the wall construction feature vector to the dimensionality of the style output via a fully connected (FC) layer, concatenates the two representations, and maps the result to the target output dimension via a second FC layer. The weights of both additional FC layers were included in the fine-tuning of the modified wall construction classifier.

Both datasets were partitioned into approximate 70% / 20% / 10% training, validation, and test splits. For the wall construction dataset, which contains multiple views per property, splits were assigned at the property level to prevent data leakage. In all cases, the test set was established at the outset, held entirely separate during fine-tuning, and used exclusively for the computation of final performance metrics. An initial training/validation split was used for hyperparameter optimisation; once an optimal hyperparameter configuration was identified, ten independent random splits were generated and used for evaluation (as described in Section 3.6). To improve image diversity and enhance the generalisation capacity of the fine-tuned models, random image transformations were applied during training. Both tasks used random cropping and horizontal flipping; for wall construction, the addition of random rotation, perspective distortion, and photometric transformations was found to improve performance further. The principal hyperparameters optimised during fine-tuning were the learning rate, and the optimiser's momentum and weight decay. All fine-tuning was performed using Stochastic Gradient Descent with a batch size of 32 and a maximum of 30 epochs. Learning rate candidates were identified through an initial learning rate range test to determine the upper bound at which loss remained stable. A cross-entropy loss was used for both tasks, as is standard for multi-class classification. However, given the pronounced class imbalance in the wall construction dataset, class weights inversely proportional to class frequency were incorporated into the loss function.

The modelling choices were intentionally conventional. The objective was not to optimise state-of-the-art performance or introduce an architectural innovation, but to establish whether a reproducible, widely used transfer-learning baseline could extract retrofit-relevant signals from accessible datasets. This design choice makes the observed generalisation failures more informative: if a standard, robust baseline performs well in-distribution but degrades on municipal samples, the result indicates a data-domain and deployment problem rather than merely a failure of an unusually complex method.

3.4. Prediction pipeline

The general approach followed four sequential steps.

Step 1: Define the list of properties of interest. This information is held by the respective councils, though with varying degrees of completeness.

Step 2: Match properties to publicly accessible data. The extent of available information determines the ease of this matching process. Many housing associations do not record the national Unique Property Reference Number (UPRN), instead using an internal identifier. The authoritative dataset linking property addresses to national UPRNs is AddressBase, maintained by Royal Mail, which is not freely available. As a workaround, the EPC register can be queried by address – as EPC certificates almost invariably contain the UPRN – which in turn enables geolocation of each property via the Open UPRN dataset maintained by Ordnance Survey, which is freely accessible.

Step 3: Collect images for each property. For the features examined in this study, street-level imagery provides adequate coverage and can be queried by geolocation. The acquisition date of each image is important, as it should correspond as closely as possible to the EPC lodgement date. While static camera parameters can be defined using a straightforward heuristic, more advanced methods exist [32] to capture multiple views per property more robustly.

Step 4: Run model inference on each collected image. Pre-classification filtering using an image segmentation model can be applied to exclude images in which building façades are insufficiently visible due to occluding elements such as trees or vehicles.

3.5. Evaluation area selection and data collection

Borough selection was carried out via convenience sampling, using Retrofit London as a channel to solicit collaboration from a broad range of local authorities. Havering and Haringey were the two boroughs that responded. The specific properties within each borough were selected on the basis of recently completed (Havering) or ongoing (Haringey) retrofit works, so that the methodology could be applied to a sample closely representative of the stock that other local authorities are likely to prioritise for retrofitting. Retrofit works in both boroughs were funded through the UK Government's Social Housing Decarbonisation Fund (SHDF). Later in the project, the Greater London Authority (GLA) expressed interest in the work and contributed an additional property sample for evaluation. These properties were drawn from an area of the London Borough of Brent selected to ensure a representative mix of wall construction types and build periods. The evaluation was restricted to houses, as these typically fall within the frame of a single street view image – a condition that cannot be reliably assumed for apartment buildings.

Fig. 5. Example images used for evaluation. ©2019 Google.

This restriction has implications for the pipeline's generalisability. Extending the method to larger multi-unit buildings would require additional steps, including façade segmentation, unit-to-building attribution, and multi-view aggregation. While these considerations are important for a robust, deployment-ready pipeline, they fall outside the scope of this pilot study. This arguably limits the generality of the observations, given that multi-unit apartment blocks form a major part of London's social housing stock by number of dwellings. However, low-rise houses remain a substantial and operationally relevant component of many local authority portfolios by number of buildings. The present study therefore addresses a bounded but meaningful subset of the stock. Moreover, the challenges surfaced by this work are not limited to single-occupancy houses and would carry over to apartment blocks.

Given the varying levels of detail available across sources, missing attributes were supplemented using the Energy Performance Certificate register for England and Wales – an open-access dataset offering both direct download and API access. Property images were sourced from GSV (see Fig. 5), which currently represents the only street-level imagery platform with sufficient spatial and temporal coverage to obtain images of the specific properties within a comparable acquisition window (typically within one to two years). Architectural style labels for the evaluation sample were assigned manually.

3.6. Quantitative evaluation strategy

The primary evaluation metrics for both classification tasks were overall accuracy and the macro-averaged F_1 score. To obtain the latter, per-class F_1 scores were computed independently and their arithmetic mean taken, an approach that ensures adequate sensitivity to minority-class performance. All metrics were averaged across ten independently fine-tuned models, each initialised with a different random seed, resulting in distinct training/validation splits and image transformation sequences per run. For each task, confusion matrices were generated using the model achieving the highest test-set F_1 score among the ten; these matrices are intended to illustrate the typical distribution of errors across classes rather than to report absolute performance.

The primary evaluation dataset was the held-out test set derived from the initial data split. To assess the applicability of the trained models to real-world deployment conditions, the property samples provided by the local authorities and the GLA (described in Section 3.2) were used as out-of-distribution (OOD) evaluation sets. OOD data refers to samples drawn from a different distribution than the training data – whether in terms of collection methodology, geographic context, or image characteristics. Performance on OOD data therefore provides an assessment of a model's generalisation capability to inputs that differ systematically from its training distribution.

To characterise the nature and magnitude of the distributional shift between the in-distribution (ID) test sets and their corresponding OOD evaluation sets, three complementary analyses were conducted at different levels of abstraction. First, a pixel-level statistical analysis compared per-channel mean and standard deviation across the two sets, providing a low-level summary of photometric differences arising from variation in capture conditions, sensor characteristics, or post-processing. Second, a frequency-domain analysis examined the radial power spectra of each dataset; the slope of a linear fit to log power against log frequency was used as a compact descriptor of the spectral profile, with differences in slope indicating systematic variation in fine-grained textural structure that is known to influence CNN feature representations [38]. Third, an embedding-space analysis quantified semantic distributional shift using Kernel Inception Distance (KID) [39]. Features were extracted using a pretrained CLIP ViT-B/32 model [40] rather than the task-specific fine-tuned models, yielding embeddings that are comparable across the two tasks. KID was preferred over Fréchet Inception Distance (FID) because the OOD evaluation sets contain fewer than one thousand samples, a regime in which FID's Gaussian approximation introduces substantial bias [41]. Together, these three analyses provide a structured account of where and how the ID and OOD sets diverge, which in turn contextualises the degradation in classification performance observed on OOD data.

Ablation experiments were not treated as a central objective because the study does not propose a new model component whose isolated effect must be demonstrated against a state-of-the-art baseline. Nevertheless, the comparison between the wall-construction model with and without the architectural-style auxiliary input provides a limited task-specific ablation of the only additional fusion component introduced in the study.

3.7. Stakeholder workshop

Towards the end of the pilot study, a workshop was convened to examine challenges surrounding building data collection and integration for retrofit archetyping. The fifteen participants included project partners (Retrofit London and the London Office for Technology and Innovation), retrofit organisations and consultancies (Energiesprong, Ambue, Domna, xRI), one architectural practice (Constructive Thinking), representatives from Havering Council and the GLA, and three members of the research team. Presentations by the authors and three external participants addressed themes central to the workshop, followed by structured group discussions. A thematic analysis was subsequently conducted based on the presentations and notes recorded during the session.

4. Results

This section first presents the performance of the fine-tuned computer vision models on their respective ID test sets (Section 4.1), before characterising the image domain shift between ID and OOD sets (Section 4.2), which provides context for the performance results on the OOD test set (Section 4.3). Performance results are summarised in Table 1.

4.1. Model performance on ID test datasets

The wall construction classifier, without style inference, achieves an accuracy of 93.8% on the ID test set but an F_1 score of 62.3%, indicating a failure to adequately learn the minority class. This is confirmed by the confusion matrix in Fig. 6, which shows that all timber frame buildings were misclassified as solid brick. Nevertheless, the model demonstrates an ability to discriminate between solid brick and cavity wall constructions where sufficient training samples were available. The architectural style classifier achieves an accuracy of 93.7% and an F_1 score of 90.7%, reflecting substantially stronger performance across minority classes. As shown in the confusion matrix in Fig. 6, misclassifications tend to occur between historically adjacent styles, suggesting that the model has acquired a meaningful latent representation of architectural chronology. The superior performance of the style classifier relative to the wall construction classifier – despite the greater number of target classes – may be attributable to a more balanced class distribution in the style dataset. An additional explanatory factor is that architectural style is an inherently visual characteristic, whereas wall construction type may be concealed beneath exterior finishes such as render or cladding. It is also notable that the model fine-tuned by Lindenthal and Johnson [30] on the same dataset achieved a macro-averaged F_1 score of 65.9%; the considerably higher performance obtained here is likely a consequence of employing a more expressive CNN architecture alongside more rigorous data augmentation and hyperparameter optimisation.

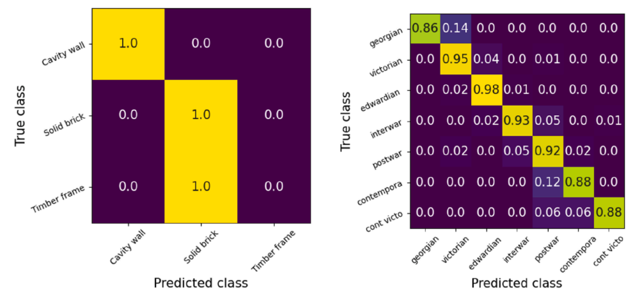

Fig. 6. Confusion matrices for wall construction (left) and architectural style (right) classification, evaluated on the in-distribution test set.

Given the correlation between architectural style, construction age, and wall construction type, an extension of the wall construction classifier was trained to investigate whether the style classifier's output could improve the standalone model's predictions. The augmented wall construction model achieves an accuracy of 96.4% and an F_1 score of 64.7%, representing a modest improvement over the baseline. However, the magnitude of this gain is limited, which may reflect the fact that the style classifier was not trained on the wall construction dataset and was applied purely in inference mode, without any task-specific fine-tuning.

4.2. Analysis of the ID-OOD image domain shift

The three analyses reveal a consistent pattern: the distributional gap between the ID test set and the OOD evaluation set is substantially larger for the architectural style task than for the wall construction task, suggesting that the two models face qualitatively different generalisation challenges (Table 2).

At the pixel level, the architectural style OOD images are markedly brighter and exhibit greater contrast than the ID images, with per-channel mean differences of approximately 30–37 points (on a 0–255 scale) and standard deviations elevated by 10–17 points across all three channels. This is consistent with the collection methodology differences noted in Section 3.6: the OOD images were captured at a wider crop and under different lighting conditions than the Cambridge training data. By contrast, the wall construction datasets are considerably closer in photometric character, with mean differences of fewer than 15 points per channel and closer standard deviations, reflecting the fact that both sets were drawn from broadly similar street-level imagery pipelines.

The frequency-domain analysis reinforces this picture. For architectural style, the ID spectral slope (−3.27) is notably steeper than the OOD slope (−2.74), indicating that the training images have a stronger high-frequency roll-off – that is, relatively less fine-grained textural

Table 1
Classification performance (accuracy and F_1) of the fine-tuned models across the in-distribution (ID) and out-of-distribution (OOD) test sets. For each task, 10 rounds of fine-tuning were run on different training/validation random splits, and the performance of each resulting model on the test sets was averaged (mean ± standard error of the mean, %).

	Accuracy		F1	
Model	ID	OOD	ID	OOD
Architectural style classifier	93.71 ± 0.19	56.41 ± 1.03	90.71 ± 0.49	27.77 ± 1.40
Wall construction classifier	93.78 ± 0.41	78.76 ± 2.12	62.25 ± 0.39	41.82 ± 2.08
Wall construction classifier with arch. style	96.35 ± 0.76	82.34 ± 2.02	64.66 ± 0.54	39.88 ± 2.98

Table 2
Pixel-level statistics, spectral slope, and Kernel Inception Distance (KID) for the in-distribution (ID) test sets and out-of-distribution (OOD) evaluation sets for each task. Pixel-level metrics are on a 0–255 scale. Spectral slope is the gradient of a linear fit to log radial power versus log frequency ($RQ1^2$ in parentheses). KID is reported as mean ± standard deviation.

		Architectural Style data		Wall Construction data	
Analysis	Metric	ID	OOD	ID	OOD
Pixel-level (channels)	Mean R	102.91	134.34	124.79	134.32
	Mean G	100.58	131.13	119.54	131.13
	Mean B	90.20	126.79	112.08	126.76
	Standard dev. R	51.69	64.62	57.33	64.59
	Standard dev. G	51.37	66.22	58.11	66.19
	Standard dev. B	54.24	71.67	62.97	71.56
Frequency-domain	Spectral slope (R^2)	−3.272 (0.934)	−2.740 (0.970)	−2.841 (0.949)	−2.724 (0.972)
Embedding-space	KID (mean ± std)	0.1750 ± 0.0086		0.0429 ± 0.0043	

energy – compared to the OOD images, which may stem from a higher level of compression. Since CNNs are known to be sensitive to the spectral statistics of their training distribution [38], this discrepancy implies that the architectural style model may have calibrated its feature responses to a different textural regime than it encounters in the OOD dataset. For wall construction, the slopes are much closer (−2.84 versus −2.72), suggesting that the textural character of the OOD images poses little additional challenge relative to training.

The embedding-space analysis corroborates both findings in a single summary statistic. The architectural style KID of 0.175 indicates a relatively large semantic distributional gap, while the wall construction KID of 0.043 is approximately four times smaller, suggesting that the OOD wall construction images occupy a region of feature space considerably closer to the training distribution. It should be noted, however, that KID operates on semantic embeddings and therefore cannot distinguish between shift attributable to image capture differences and shift arising from genuine variation in the underlying subject matter. For architectural style in particular, Cambridge and London may differ sufficiently in their built environment (e.g. in terms of vernacular style variations or building materials) that some portion of the measured KID reflects a real architectural difference rather than a purely photometric or methodological artefact. The pixel-level and spectral analyses, which are insensitive to semantic content, provide a useful counterpart in this respect. Taken together, the three analyses establish that the architectural style model is exposed to a more severe and multi-faceted domain shift at evaluation than the wall construction model, a distinction that should be borne in mind when interpreting the OOD classification results that follow.

4.3. Model generalisation to OOD test data

The wall construction classifier without style inference achieves an accuracy of 78.8% and an F_1 score of 41.8% on the OOD dataset, underscoring the difficulty of learning minority classes under distribution shift. As illustrated in Fig. 7, despite the absence of timber frame buildings in the OOD set, performance remains low due to a systematic tendency to classify cavity wall samples as the dominant class, solid brick.

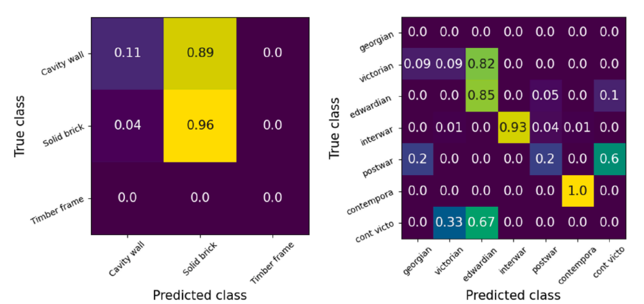

Fig. 7. Confusion matrices for wall construction (left) and architectural style (right) classification, evaluated on the OOD set (borough and GLA data).

The architectural style classifier exhibits even more severe generalisation degradation, achieving an accuracy of 56.4% and an F_1 score of 27.8%. The second confusion matrix in Fig. 7 further reveals a susceptibility to revival styles, with contemporary Victorian-style properties frequently misclassified as Victorian or Edwardian.

Given the limited generalisation capacity of the style classifier, its contribution to wall construction prediction in the OOD setting is ambiguous. Although the combined model achieves an improved accuracy of 82.3%, and its F_1 score deteriorates to 39.9%, the higher standard error across the 10 runs makes it difficult to draw a clear conclusion. This outcome highlights a fundamental limitation of leveraging publicly accessible computer vision datasets not purpose-built for retrofit applications: even where such datasets are large and exhibit favourable class distributions, the source image domain may diverge sufficiently from the deployment context to impair practical utility.

4.4. Mapping example: construction age and architectural style

Retrofit archetypes serve to group properties on the premise that greater similarity in physical characteristics implies greater similarity in the interventions required. Efficient identification and mapping of such property groups is therefore a prerequisite for delivering retrofit at scale.

Fig. 8. EPC-derived construction age for a sample of buildings in the OOD dataset (top) and a street-level image of the properties in the green rectangle (bottom). "Ambiguous" denotes buildings in which multiple units were assigned different construction ages. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

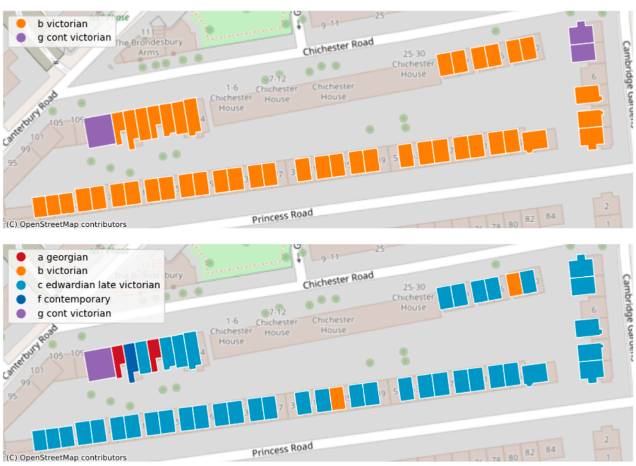

Fig. 9. Architectural style of a sample of buildings in the OOD dataset (top: manual labels; bottom: predictions from the fine-tuned classifier).

Beyond predictive accuracy, an important consideration in this context is the consistency of model predictions: how reliably does a computer vision model assign the same label to visually similar buildings, and how does this compare to the consistency of equivalent attributes recorded in public registers? To address this question, this section goes back to the relationship between construction age and architectural style. EPC data are known to exhibit inconsistencies arising from inter-assessor disagreement on building parameters [8]. As shown in Fig. 8, construction age can vary substantially along a single street of visually homogeneous properties, or even within a single multi-unit building. In extreme cases, adjacent properties may be assigned to age bands separated by several decades – as illustrated by the orange footprint (1899–1929) appearing immediately adjacent to the blue footprint (1949–1966).

Architectural style, as a visually grounded characteristic, is more amenable to consistent assignment across similar-looking properties, as demonstrated in Fig. 9. This consistency is partially preserved in the predictions of the fine-tuned classifier – though shifted toward historically proximate style categories, reflecting the generalisation challenges discussed in Section 4.3. Despite this limitation, the spatial coherence of the model's outputs suggests meaningful potential as a decision-support tool for retrofit archetyping at scale.

4.5. Stakeholder insights on data infrastructure

The above results were presented during the stakeholder workshop described in Section 3.7. Based on the group discussions that followed, four main themes emerged.

Fragmented data governance and barriers to access. Participants consistently highlighted the fragmented landscape of retrofit-related data governance. Issues included inconsistent data custodianship, limited access to essential datasets, and commercial incentives that restrict data sharing. Local authorities noted a distinction between having data and having confidence in its reliability, underscoring gaps in quality assurance and validation. The need for staged data-infrastructure development was also emphasised, reflecting the difficulty of coordinating data flows across diverse stakeholders. These access barriers had direct methodological consequences: the limited geographic scope of the available training sets is a plausible contributor to the sharp OOD performance drops observed – particularly the decline in architectural style F_1 from 90.7% in-distribution to 27.8% on borough data – since locally representative training samples could not be assembled within the constraints of the pilot.

Standardisation as a foundation for effective archetyping. A major theme concerned the lack of standardised archetypes and the broader challenge of data standardisation. Participants stressed that archetyping accuracy depends on shared definitions of property characteristics and more granular data (e.g., insulation quality rather than binary categories). Agreement on common standards was seen as essential for interoperability, comparability, and meaningful integration of datasets. Standardisation was also linked to the feasibility of automated assessments and AI-driven processes. The concrete cost of this inconsistency is visible in the quantitative results – for instance, in the construction age map in Fig. 8, adjacent properties are assigned to age bands decades apart, illustrating assessor disagreement and a lack of spatial coherence.

Balancing automation and human oversight in assessment. The workshop explored the role of automation in property assessment and retrofit planning. A three-level model was discussed, ranging from automated anomaly detection to human validation and on-site verification. While automation was viewed as valuable – especially for high-level, strategic decision-making – participants insisted on maintaining a "human in the loop," both to ensure data quality and to retain professional accountability. This balance was seen as crucial for trustworthiness and operational practicality. The wall construction and architectural style classifier's drop in performance on borough data – despite performing well under benchmark conditions – is a direct illustration of the need for human oversight in a deployment context.

The need for participatory and hybrid (top-down and bottom-up) data approaches. Finally, participants discussed the importance of complementary top-down and bottom-up data strategies. While current archetyping is largely top-down, there was recognition of the value in bottom-up approaches, including resident participation to correct or enhance datasets, and exploratory data-driven techniques such as unsupervised machine learning (e.g., clustering). Such hybrid models were viewed as necessary for building robust, scalable, and locally accurate retrofit archetypes.

5. Discussion

This study aims to determine whether publicly available housing data and computer vision techniques could be used in supporting retrofit planning in social housing. The findings of this study indicate that there is indeed a clear but qualified role for using publicly available housing data and computer vision in supporting retrofit planning in social housing.

The first key finding of this study is that there is a distinction between task suitability and data suitability. The study found that the architectural style classifier had good performance in-distribution since architectural styles are visually salient features. However, the study also found that the wall construction classifier was more difficult, especially for minority classes. This shows that not all features that could be used in retrofit planning are equally well-suited for visual inference. From a planning point of view, what this means is that computer vision is likely to be most effective if used selectively on features where there is strong visual signal and where classification is not too precise.

A second important result relates to the portability of the classifier models that were developed based on publicly accessible data. Although the two classifiers were found to perform well on the respective test data sets, the performance was significantly reduced when the data was drawn from the borough and GLA data sets. The most striking case is that of architectural styles, in which the classifier that was developed based on the Cambridge data set failed to generalise well to the London data. This underlines the significance of the provenance of the data in terms of geography, vision, and methodology in the evaluation of the usefulness of the classifier. The implication is that publicly accessible data sets, even if they are large and well-labelled, may not be appropriate without closer alignment.

A third observation arises when considering the interaction with the tasks and the outputs of the models. The marginal improvement in the wall construction task when including the architectural style prediction indicates that leveraging the relationships between building features is possible but to a limited extent with the current approach. The absence of improvement in the out-of-distribution setting also indicates that relationships between features are sensitive to domain changes. This observation raises the question of the overall approach to combining multiple weak or context-dependent signals and the extent to which this can be leveraged for robust prediction. In the context of retrofitting, it suggests that multi-modal and multi-feature approaches will be important.

Other than that, the findings also point to the value of consistency in model outputs. The mapping exercise also reveals that even though there is reduced accuracy in architectural style predictions, there is spatial coherence in model outputs for similar buildings, unlike the inconsistency in construction age data from EPC. This reveals that there is value in using computer vision not just in terms of accuracy but also in terms of generating more consistent representations of the building stock. This is significant in retrofit planning, where archetyping is done on the basis of similar properties.

However, the results of the stakeholder workshop also place these technical results in the context of the realities of retrofit delivery. The results of the workshop indicated that data fragmentation, lack of standardisation, and interoperability issues are significant challenges. The role of computer vision in these circumstances is not to replace the data systems but to supplement them in areas of deficiency and provide additional information. However, the effectiveness of computer vision results is also dependent upon their integration into larger workflows. The emphasis given in the results of the workshop to the importance of having a “human in the loop” also supports the importance of results being interpretable and verifiable.

Overall, the results point towards an understanding of the role of AI in retrofit planning that is centred on supporting evidence generation in situations of uncertainty. Rather than providing clear classifications, models were seen to be effective in identifying potential building characteristics, pinpointing inconsistencies, and supporting the development of initial archetypes. This is particularly relevant in early-stage planning where there is often a need for decision-making on housing stocks with limited information. Here, the ability to provide consistent, scalable, and cost-effective approximations is more significant than seeking accuracy.

Finally, the findings point toward a more incremental and socio-technical approach toward smart retrofit systems. This is because the integration of model outputs, publicly available data, local authority data, and knowledge from stakeholders indicates that effective retrofit intelligence is created by integrating multiple data sources. Therefore, computer vision is likely part of this system, but its effectiveness is subject to alignment with data standards, governance, and decision-making processes. Therefore, in order to make further progress in this space, it is not simply a question of improving models but also creating more collaborative data infrastructures.

The practical implication is that the proposed pipeline should be viewed as a research prototype for evidence generation, not as a procurement-ready tool. Its immediate use is to reveal where existing data are sufficient, where they fail, and what labelled datasets would be required for responsible scale-up.

6. Limitations and future research

6.1. Data quality and representativeness

A recurring constraint throughout this study is the quality and representativeness of the available training data. The wall construction dataset exhibits a pronounced class imbalance which likely constrains the model’s ability to reliably identify minority classes, even with inverse-frequency class weighting applied during training. This imbalance is reflected directly in the results: despite high overall accuracy, the F_1 scores for both the in-distribution and OOD evaluations reveal systematic misclassification of underrepresented classes. The dataset’s geographic restriction to London further limits its representativeness of the broader UK housing stock.

The architectural style dataset presents complementary limitations. Only 6345 of the approximately 25,000 labelled Cambridge images from Lindenthal and Johnson [30] were recoverable, and it is unclear whether this subset is representative of the full class distribution. More consequentially, the geographic provenance of this dataset (i.e., Cambridge) diverges from the London deployment context, and this domain gap is evidenced by the sharp decline in style classifier performance on the OOD evaluation set, from an F_1 score of 90.7% in-distribution to 27.8% on borough data. The use of EPC records as a source of wall construction labels inherits well-documented quality issues, even if social housing EPCs are generally considered more reliable than those for privately owned stock. Label noise remains a plausible contributor to observed performance gaps but is difficult to disentangle from model or image quality effects. A dedicated investigation – comparing model predictions against ground-truth labels derived from expert surveys on a subset of properties – would help isolate this contribution, particularly in the OOD setting. SVI introduces additional confounders, including variable acquisition dates, occlusion by vegetation, vehicles or street furniture, and inconsistent camera angles. These factors complicate both reproducibility and scaling.

The study also depends on data that are publicly accessible but not uniformly open. In particular, GSV is subject to platform licensing and redistribution restrictions. This does not prevent reproducibility of the method, since the same API-based acquisition procedure and geolocation inputs can be used by other researchers with appropriate access, but it does limit direct redistribution of the raw images. Future work should investigate the use of alternative street-level imagery sources where licensing terms are more permissive, while recognising that coverage, image quality and temporal consistency may be weaker than GSV in many UK contexts.

6.2. Modelling choices and evaluation design

The study employs ResNet-50 as its base architecture – a well-established choice, but one that predates more recent advances in computer vision. Benchmarking it against alternative architectures such as Vision Transformers or foundation models with self-supervised pre-training would help determine whether the performance ceiling reported here reflects an inherent task difficulty or a constraint of the chosen architecture. The style-wall construction fusion module, which concatenates frozen style classifier outputs with wall construction features via two fully connected layers, is similarly underexplored. The relatively modest improvement in F_1 score attributable to style inference (from 62.3% to 64.7% in-distribution, and a deterioration in the OOD setting) suggests that the current architecture does not fully exploit the correlation between style and construction type. Alternative fusion strategies could be considered, such as jointly fine-tuning the style classifier, which could increase the representational capacity of the combined model.

6.3. Scope and generalisability

The study evaluation is deliberately restricted to houses, excluding flats and apartment buildings. Given that a substantial proportion of London's social housing stock consists of medium- and high-rise blocks, this is a significant constraint on the practical generalisability of the pipeline. Larger multi-unit buildings are likely to introduce additional sources of error because a single street-level image may contain multiple entrances, multiple dwelling units, mixed construction phases, obscured elevations or façade refurbishments. Future research on multi-unit stock should therefore treat building-level segmentation and unit-level attribution as core methodological challenges. Extending the methodology to apartment buildings represents a natural and important direction for future development.

Similarly, the geographic restriction to London means that transferability to other urban contexts, rural building stock, or different national building traditions remains undemonstrated. Only two building characteristics – wall construction type and architectural style – are addressed. Many features central to retrofit archetyping, including roof construction, floor type, glazing specification, and heating system, fall outside the scope of the present study. Furthermore, the proxy relationship between architectural style and construction age – which underpins part of the rationale for style classification – is invoked but not empirically quantified within this study. Directly validating this relationship using large-scale EPC-address-image linkages, which the data pipeline developed here is well-positioned to support, would strengthen the interpretive value of style prediction as a retrofit-relevant signal.

6.4. Integration into retrofit practice

The pipeline is designed as a component that could feed into retrofit archetyping tools, but the study does not address how model outputs should be integrated into downstream decision-making. Questions of practical importance – such as how prediction uncertainty should be communicated to assessors, at what confidence threshold a prediction should trigger a manual inspection, or how outputs should be reconciled with existing EPC data – are not resolved here. Future work should examine the appropriate integration of model outputs into PAS 2035-compliant workflows, including the thresholds and override mechanisms that would be necessary for responsible deployment. Engagement with retrofit practitioners throughout this process would be essential to ensure that the tool supports, rather than supplants, professional judgement.

The stakeholder workshop indicates three priorities for such integration. First, model outputs should be presented with uncertainty and provenance information so that users can judge when to trust them. Second, automated predictions should be linked to existing records in a way that highlights disagreement rather than concealing it. Third, any operational system should include a route for human validation, including surveyor review and, where appropriate, resident feedback. These requirements are as important as model accuracy for responsible use in retrofit planning.

7. Conclusions

This paper has sought to explore the question of whether publicly accessible property data, street-level imagery, and computer vision can be harnessed to inform retrofit planning in social housing. The evidence suggests that they can be, but in a very specific and circumscribed way. The most important contribution that these sources make is likely to be in the generation of evidence in the early stages of planning for retrofitting large numbers of buildings in which information is either lacking, incomplete, or hard to integrate. The three contributions that this research makes are that it proves that publicly accessible data can be aggregated into a viable pipeline to generate evidence about building characteristics that might be of relevance to retrofitting at the property level. It proves that computer vision can perform well under benchmark conditions, particularly in the case of visually salient characteristics such as architectural style. It proves that the most important barrier to the application of these methods in practice is that they need to be translated into the context of borough and GLA datasets.

One of the most interesting findings is that performance alone is not an adequate proxy of usefulness. The large drop in performance when tested against real-world data from municipalities emphasizes the importance of generalization, data provenance, and context of use. However, the results also demonstrate that there is potential in consistency, scalability, and low-cost approximations, particularly when the goal is not classification but rather archetyping, screening, and prioritization. From this perspective, the value of AI is best characterized as decision support in the face of uncertainty. Another interesting finding is that the usefulness of these approaches is not only dependent on the quality of the models but also on the data infrastructure and management practices. The discussions with stakeholders highlighted the problem of fragmented data infrastructure, lack of consistency in standards, and interoperability as significant obstacles to effective retrofit planning. This points to the conclusion that the future of smart retrofit planning will not only be dependent on improved models but also on more integrated, transparent, and collaborative data management practices. Computer vision can be an aid in more informative data environments but is not a replacement for poor integration of data management practices.

On a more general level, the results support the socio-technical perspective that is taken in the concept of smart retrofit systems. Indeed, effective retrofit intelligence is likely to be the outcome of the integration of automated inference, records, practitioner expertise, and, where relevant, resident understanding. With regard to the potential of AI in this context, the key point is that AI is likely to be supportive but not autonomous. In conclusion, the findings of this study imply that publicly available data and computer vision do hold potential in the context of social housing retrofit planning, particularly in the early stages of the stock. However, the potential of publicly available data and computer vision is likely to be most relevant in the near term in supporting the transition of organisations from fragmented evidence towards more cohesive forms of evidence. To do so, however, is likely to require advances not only in machine learning but also in standardisation, the development of appropriate labelled datasets, and collaborative approaches to the use of digital tools.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used Anthropic's Claude in order to refine the flow and tone of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

CRediT authorship contribution statement

Robin Roussel: Methodology, Software, Conceptualization; **Luke Gooding:** Writing – original draft, Data curation.

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.

Acknowledgements

This work was supported by the Ecological Citizen(s) Network+, a four-year research network based at the Royal College of Art in collaboration with Wrexham University and the Stockholm Environment Institute at the University of York, funded by the EPSRC Network+ award (EP/W020610/1). The authors wish to thank Prof Sam Jacoby (Royal College of Art) for his supervision and contributions to funding acquisition and methodological development. We are grateful to Jack Ostrofsky at Warmer Homes London (previously Retrofit London) and Jay Saggar at the London Office of Technology and Innovation for their invaluable assistance in establishing key contacts for evaluation data collection and stakeholder engagement. We also thank the Councils of the London Boroughs of Havering and Haringey, ECD Architects, and the Greater London Authority for providing evaluation data, and all participants in the stakeholder workshop for their contributions and insights.

References

[1] UNEP, GlobalABC, Not Just Another Brick in the Wall: the Solutions Exist - Scaling them will Build on Progress and cut Emissions Fast. Global Status Report For Buildings and Construction 2024/2025, United Nations Environment Programme, 2025. https://doi.org/10.59117/20.500.11822/47214

[2] HBF, Housing Horizons: Examining UK Housing Stock in an International Context, Technical Report, Home Builders Federation, 2023. https://www.hbf.co.uk/news/housing-horizons/.

[3] MHCLG, English Housing Survey 2021 to 2022: Headline Report, Technical Report, Ministry of Housing, Communities & Local Government, 2022.

[4] P. Brown, W. Swan, S. Chahal, Retrofitting social housing: reflections by tenants on adopting and living with retrofit technology, Energy Effic. 7 (4) (2014) 641–653. https://doi.org/10.1007/s12053-013-9245-3

[5] G. Killip, A. Owen, E. Morgan, M. Topouzi, A co-evolutionary approach to understanding construction industry innovation in renovation practices for low-carbon outcomes, Int. J. Entrep. Innov. 19 (1) (2018) 9–20. https://doi.org/10.1177/1465750317753933

[6] R. Kitchin, The Data Revolution: Big Data, Open Data, Data Infrastructures & their Consequences, SAGE Publications Ltd, 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom, 2014. https://doi.org/10.4135/9781473909472

[7] I. Hamilton, A. Summerfield, T. Oreszczyn, P. Ruyssevelt, Using epidemiological methods in energy and buildings research to achieve carbon emission targets, Energy Build. 154 (2017) 188–197. https://doi.org/10.1016/j.enbuild.2017.08.079

[8] 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

[9] MHCLG, Energy Performance of Buildings Certificates Statistical Release: January to March 2026, England and Wales, Technical Report, Ministry of Housing, Communities & Local Government, 2026. https://www.gov.uk/government/statistics/energy-performance-of-building-certificates-in-england-and-wales-january-to-march-2026/energy-performance-of-buildings-certificates-statistical-release-january-to-march-2026-england-and-wales.

[10] P. De Wilde, The gap between predicted and measured energy performance of buildings: a framework for investigation, Autom. Constr. 41 (2014) 40–49. https://doi.org/10.1016/j.autcon.2014.02.009

[11] P. Steadman, S. Evans, R. Liddiard, D. Godoy-Shimizu, P. Ruyssevelt, D. Humphrey, Building stock energy modelling in the UK: the 3DStock method and the London Building Stock Model, Build. Citi. 1 (1) (2020) 100–119. https://doi.org/10.5334/bc.52

[12] T. Mac Uidhir, F. Rogan, M. Collins, J. Curtis, Ó.G. Brian P., Improving energy savings from a residential retrofit policy: a new model to inform better retrofit decisions, Energy Build. 209 (2020) 109656. https://doi.org/10.1016/j.enbuild.2019.109656

[13] R. Hecht, T. Danke, H. Herold, P. Hudson, M. Munke, T. Rieche, Colouring cities: a citizen science platform for knowledge production on the building stock - potentials for urban and architectural history, in: S. Münster, A. Pattee, C. Kröber, F. Niebling (Eds.), Research and Education in Urban History in the Age of Digital Libraries, 1853, Springer Nature Switzerland, Cham, 2023, pp. 145–164. https://doi.org/10.1007/978-3-031-38871-2_9

[14] M.B. Starzyńska-Grześ, R. Roussel, S. Jacoby, A. Asadipour, Computer vision-based analysis of buildings and built environments: a systematic review of current approaches, ACM Comput. Surv. 55 (13s) (2023) 1–25. https://doi.org/10.1145/3578552

[15] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places: a 10 million image database for scene recognition, IEEE Trans. Pattern Anal. Mach. Intell. 40 (6) (2018) 1452–1464. https://doi.org/10.1109/TPAMI.2017.2723009

[16] F. Biljecki, K. Ito, Street view imagery in urban analytics and GIS: a review, Landsc. Urban Plan. 215 (2021) 104217. https://doi.org/10.1016/j.landurbplan.2021.104217

[17] H. Xu, H. Sun, L. Wang, X. Yu, T. Li, Urban architectural style recognition and dataset construction method under deep learning of street view images: a case study of Wuhan, ISPRS Int. J. Geo-Inf. 12 (7) (2023) 264. https://doi.org/10.3390/ijgi12070264

[18] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA, 2016, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90

[19] C.F. Reinhart, C. Cerezo Davila, Urban building energy modeling - a review of a nascent field, Build. Environ. 97 (2016) 196–202. https://doi.org/10.1016/j.buildenv.2015.12.001

[20] C. Wilson, L. Crane, G. Chryssochoidis, Why do homeowners renovate energy efficiently? Contrasting perspectives and implications for policy, Energy Res. Soc. Sci. 7 (2015) 12–22. https://doi.org/10.1016/j.erss.2015.03.002

[21] IEA, Net Zero Roadmap: A Global Pathway to Keep the 1.5 °C Goal in Reach, Technical Report, International Energy Agency, 2023. https://www.iea.org/reports/net-zero-roadmap-a-global-pathway-to-keep-the-15-c-goal-in-reach.

[22] B.K. Sovacool, S. Griffiths, Culture and low-carbon energy transitions, Nat. Sustain. 3 (9) (2020) 685–693. https://doi.org/10.1038/s41893-020-0519-4

[23] D. Majcen, L. Itard, H. Visscher, Statistical model of the heating prediction gap in Dutch dwellings: relative importance of building, household and behavioural characteristics, Energy Build. 105 (2015) 43–59. https://doi.org/10.1016/j.enbuild.2015.07.009

[24] BRE, RdSAP10 Specification, Technical Report, Building Research Establishment, 2025.

[25] L.G. Swan, V.I. Ugursal, Modeling of end-use energy consumption in the residential sector: a review of modeling techniques, Renew. Sustain. Energy Rev. 13 (8) (2009) 1819–1835. https://doi.org/10.1016/j.rser.2008.09.033

[26] S. Nejadshamsi, U. Eicker, J. Bentahar, C. Wang, Improving urban-scale building occupancy and energy use estimation using a transportation-informed building occupancy estimation framework, Energy Build. 333 (2025) 115468. https://doi.org/10.1016/j.enbuild.2025.115468

[27] C. Beltrán-Velamazán, M. Monzón-Chavarrías, B. López-Mesa, Predicting energy and emissions in residential building stocks: national UBEM with energy performance certificates and artificial intelligence, Appl. Sci. 15 (2) (2025) 514. https://doi.org/10.3390/app15020514

[28] M. Batty, Digital twins, Environ. Plan. B: Urban Anal. City Sci. 45 (5) (2018) 817–820. https://doi.org/10.1177/2399808318796416

[29] J. Von Platten, C. Sandels, K. Jörgensson, V. Karlsson, M. Mangold, K. Mjörnell, Using machine learning to enrich building databases-methods for tailored energy retrofits, Energies 13 (10) (2020) 2574. https://doi.org/10.3390/en13102574

[30] T. Lindenthal, E.B. Johnson, Machine learning, architectural styles and property values, J. Real Estate Finance Econ. 71 (3) (2025) 353–384. https://doi.org/10.1007/s11146-021-09845-1

[31] M. Dai, W.O.C. Ward, G. Meyers, D. Densley Tingley, M. Mayfield, Residential building facade segmentation in the urban environment, Build. Environ. 199 (2021) 107921. https://doi.org/10.1016/j.buildenv.2021.107921

[32] R. Roussel, S. Jacoby, A. Asadipour, Robust building identification from street views using deep convolutional neural networks, Buildings 14 (3) (2024) 578. https://doi.org/10.3390/buildings14030578

[33] X. Liang, W. Yap, F. Biljecki, Heterogeneous graph neural networks for building attribute prediction from hierarchical urban features and cross-view imagery, ISPRS J. Photogramm. Remote Sens. 234 (2026) 185–204. https://doi.org/10.1016/j.isprsjprs.2026.02.016

[34] M. Rosenfelder, M. Wussow, G. Gust, R. Cremades, D. Neumann, Predicting residential electricity consumption using aerial and street view images, Appl. Energy 301 (2021) 117407. https://doi.org/10.1016/j.apenergy.2021.117407

[35] K. Mayer, L. Haas, T. Huang, J. Bernabé-Moreno, R. Rajagopal, M. Fischer, Estimating building energy efficiency from street view imagery, aerial imagery, and land surface temperature data, Appl. Energy 333 (2023) 120542. https://doi.org/10.1016/j.apenergy.2022.120542

[36] M. Sun, R. Bardhan, Identifying hard-to-decarbonize houses from multi-source data in Cambridge, UK, Sustain. Citi. Soc. 100 (2024) 105015. https://doi.org/10.1016/j.scs.2023.105015

[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Miami, FL, 2009, pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[38] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F.A. Wichmann, W. Brendel, ImageNet-trained CNNs are Biased Towards Texture; Increasing Shape Bias Improves Accuracy and Robustness, 2018. https://doi.org/10.48550/ARXIV.1811.12231

[39] M. Bińkowski, D.J. Sutherland, M. Arbel, A. Gretton, Demystifying MMD GANs, 2018. https://doi.org/10.48550/ARXIV.1801.01401
[40] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: Proceedings of the 38th International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763. https://proceedings.mlr.press/v139/radford21a.html.
[41] M.J. Chong, D. Forsyth, Effectively unbiased FID and inception score and where to find them, 2020, pp. 6070–6079. https://openaccess.thecvf.com/content_CVPR_2020/html/Chong_Effectively_Unbiased_FID_and_Inception_Score_and_Where_to_Find_CVPR_2020_paper.html.