Computer vision-based analysis of buildings and built environments: A systematic review of current approaches

MAŁGORZATA B. STARZYŃSKA-GRZEŚ*, ROBIN ROUSSEL*†, and SAM JACOBY, Laboratory for Design and Machine Learning, Royal College of Art, United Kingdom
ALI ASADIPOUR, Computer Science Research Centre, Royal College of Art, United Kingdom

Analysing 88 sources published from 2011 to 2021, this paper presents a first systematic review of the computer vision-based analysis of buildings and the built environment. Its aim is to assess the potential of this research for architectural studies and the implications of a shift to a crossdisciplinarity approach between architecture and computer science for research problems, aims, processes, and applications. To this end, the types of algorithms and data sources used in the reviewed studies are discussed in respect to architectural applications such as a building classification, detail classification, qualitative environmental analysis, building condition survey, and building value estimation. Based on this, current research gaps and trends are identified, with two main research aims emerging. First, studies that use or optimise computer vision methods to automate time-consuming, labour-intensive, or complex tasks when analysing architectural image data. Second, work that explores the methodological benefits of machine learning approaches to overcome limitations of conventional analysis in order to investigate new questions about the built environment by finding patterns and relationships between visual, statistical, and qualitative data. The growing body of research offers new methods to architectural and design studies, with the paper identifying future challenges and directions of research.

CCS Concepts: • **General and reference → Surveys and overviews**.

Additional Key Words and Phrases: architecture, built environment, computer vision, machine learning, image data

1 INTRODUCTION: COMPUTER VISION IN BUILT ENVIRONMENT STUDIES

A growing number of disciplines, including architecture, are adopting data-driven applications to process large digital datasets in support of analytical and decision-making processes [21]. This paper reviews the use of images of the built environment in computer vision studies and how in turn computer vision methods find increasing use in built environment studies. This shift has important implications for research problems, aims, processes, and applications as well as terminology that require clarification for design-led research at the intersection of computer science and architecture. This review assesses the methods, techniques, and data used in recent research in order to group methods and applications and identify new research directions and knowledge gaps of interest to interdisciplinary researchers.

*Corresponding Author
†Also with, Computer Science Research Centre.

Authors' addresses: Małgorzata B. Starzyńska-Grześ, m.starzynska@network.rca.ac.uk; Robin Roussel, robin.roussel@rca.ac.uk; Sam Jacoby, sam.jacoby@rca.ac.uk, Laboratory for Design and Machine Learning, Royal College of Art, School of Architecture, Kensington Gore, South Kensington, London, United Kingdom, SW7 2EU; Ali Asadipour, Computer Science Research Centre, Royal College of Art, Research Centres, Hester Road, Battersea, London, United Kingdom, SW11 4AN, ali.asadipour@rca.ac.uk.

0360-0300/2023/1-ART $15.00
https://doi.org/10.1145/3578552

The use of analog and digital data in architectural practice and theory is well established in studies of design processes [22], buildings, and urban fabrics. Common topics include manufacturing [8], design sustainability [81], environmental impact [47], and morphology [33]. Greater availability of digital data repositories such as Energy Performance Certificates (EPCs) or property- and planning-related records however creates new applications and potential to analyse the built environment. In particular, large-scale image data processing and acquisition are an emerging area of research in the built environment and studies of its design.

Computer vision methods (including image-based machine learning) applied to buildings as well as larger architectural and urban domains can be grouped into four clusters of research: (i) landmark and place recognition, (ii) generative design and modelling, (iii) remote sensing, and (iv) the analysis of urban environments. Landmark recognition approaches have been reviewed by T. Chen et al. [14] and Bhattacharya and Gavrilova [9], while Garg et al. [31] compared visual place recognition methods. New applications of artificial design and urban environment modelling were assessed by Sönmez [89] and Feng et al. [28]. A review of deep learning applications in remote sensing by Ma et al. [68] proposed a taxonomy based on four main tasks: image preprocessing, classification, change detection, and accuracy assessment. Lastly, four reviews investigated the use of street-view imagery for the analysis of urban environments [10, 17, 38, 51].

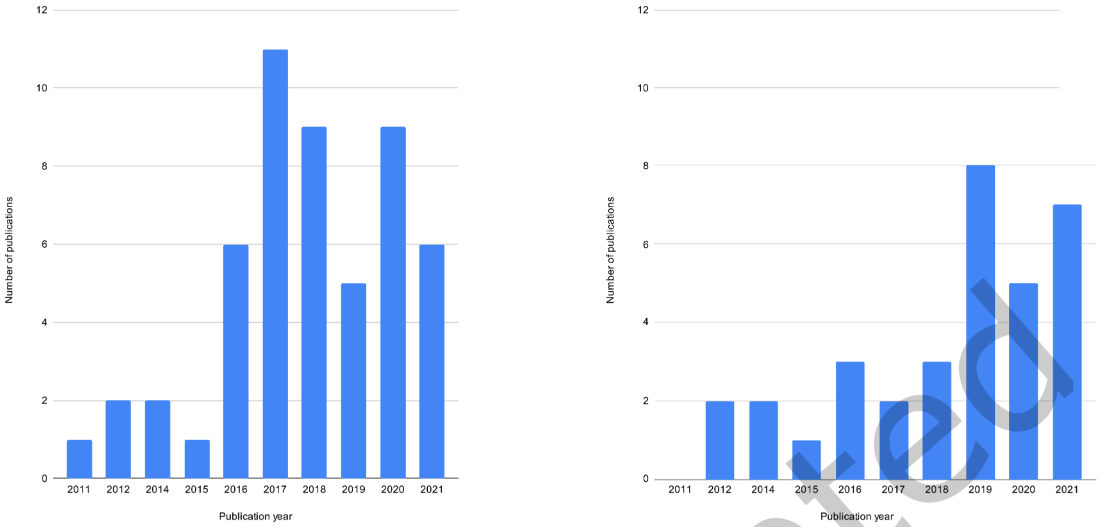
Although many of these recent reviews touch on different aspects of building recognition, there is no specific assessment of how computer vision-based applications might benefit architectural studies or takes into account architectural practices. This paper is a first systematic review of the state-of-the-art of computer vision in the analysis of the built environment, especially in relation to applications at different architectural and urban scales. It compares the research foci, computer vision and machine learning approaches, and data acquisition and curation processes found in recent studies to identify trends and challenges of this often transdisciplinary research as well as future directions and value this might bring to architectural and urban design studies. This is also of relevance to computer scientists, as the shift in focus changes the data pipeline and research approaches.

A detailed review of 88 sources identified two primary objectives in recent research: 64% of studies test or improve the performance of existing and novel algorithms by applying them to architectural datasets (Fig. 1a) and 36% assess the methodological benefits and outcomes of using computer vision techniques to ask new questions in the architectural domain (Fig. 1b). For example, the automation of architectural recognition and classification tasks can expedite otherwise labour- and time-intensive processes such as the recognition of building elements, assigning street views to specific cities or inferring neighbourhood statistics. Current research demonstrates the value of computer vision-based methods of analysis in the architectural domain, such as a correlation of visual and statistical or demographic data.

This paper explores how new questions about urban gentrification, real-estate values or specific characteristics of the built environment can be asked using computer vision methods, and how this might inform decision-making by designers, occupants, and policymakers. The aim is further to contribute to a much-needed transdisciplinary evaluation of common interests between computer vision and architectural or urban design studies to strengthen the reliability and methodological rigour of research. Disciplinary differences in understanding and assessing computer vision or spatial design problems can lead to misunderstandings that need be resolved to fully realise the potential of computer vision approaches in architecture.

This review makes the following contributions:

- It provides an application-centred classification of current research at the intersection of computer vision and architecture.
- It analyses the trends and gaps emerging from the research.
- It assesses the current state of data sources, both in terms of acquisition methods and geographic locations.
- It evaluates the reproducibility and comparability of computer vision-driven architectural research and summarises common problems and possible solutions.



(a)

(b)

Fig. 1. Publications focused on: (a) innovating machine learning tool by applying them to architectural datasets by year, (b) formulating novel questions by employing machine learning techniques in the architectural domain by year.

- It outlines opportunities for further interdisciplinary architecture and computer science research.

This review is organised into four sections. A methods section provides details on the criteria formation for the inclusion of reviewed papers and sources. This is followed by a summary of the search results and main findings. A discussion section then highlights the main research trends, challenges, and pitfalls. Finally, the concluding section makes recommendations on key future research directions.

2 METHODS

This paper uses the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) method [79, 80] and a multi-stage selection process. Primary and peer-reviewed sources were selected using a keyword search of the IEEE Xplore, JSTOR, Scopus, Semantic Scholar, and Google Scholar databases, as well as via Google Search, to identify studies using computer vision in the context of the built environment. Given a noticeable increase in studies over the last five years, this review is limited to research published from 2011 to 2021 to capture the most recent trends.

The scope of this paper is also limited to studies in architecture and urban design and excludes work from associated fields such as engineering, building technology, and interior and landscape design. In particular, related research on building interiors and 3D reconstruction was excluded, as this review focuses on research and methods with direct implications for architectural practice and design at the building scale. While acknowledging that computer vision methods are also applied to studies of interior building environments, especially in the context of plan layout generation and style recognition, this review focuses on outdoor features as interior environments lack public data sources and tend to see frequent and significant change, making them difficult to compare or analyse at scale. Likewise, this review does not include the 3D reconstruction of urban environment, as from an architectural practice perspective this is of secondary interest. This type of research tends to focus on speculative design rather than actual architectural and urban conditions, although some applications in urban planning are perceivable.

The overlap of terms widely used in both computer science and architecture became an unexpected challenge. The meaning of keywords such as “architecture”, “structure”, “model”, “design” or “building” depends on disciplinary contexts and can refer to significantly different concepts. For example, the initial search for the keywords “machine learning” together with “architecture” brought up publications that used the term “architecture” to describe the structure of machine learning systems. Using less ambiguous words such as “façade”, “urban”, “city” or “ornament” produced better results in identifying literature relevant to built environment studies. However, it created a risk of excluding papers applying computer vision methods to other aspects of architecture. This problem was mitigated using a multi-stage selection process (Fig. 2). Following the keyword search, paper titles, and abstracts were screened with respect to both computer vision and architectural or urban analysis. References in the most relevant papers were also checked. In addition, conference papers (if not published in a journal), online sources (e.g. research project websites), and PhD theses were reviewed and added. This created 226 relevant records. From these, duplicates, papers not peer reviewed or cited in peer-reviewed journals, and conference papers later republished in journals (31 records) as well as papers referring to the “architecture” or “structure” of computational systems rather than buildings (69 records) were removed. 3 reports could not be retrieved. This left 123 sources that fully met the initial search criteria.

A large body of work looked at the application of machine learning in scene recognition for the purpose of navigation and obstacle detection in self-driving vehicles. Although some relate to building analysis, due to similar methodology with papers included in this review, these were omitted from the analysis. Furthermore, studies that explored generative rather than analytical systems were excluded, as were those that only used non-visual data or 3D datasets such as point clouds. Both were considered outside the scope of the paper. Removing 35 records, finally left 88 sources for further analysis in this review.

Two types of information in the publications were compared: (i) the algorithms and methods used in relation to machine learning models and computer vision tasks and (ii) architectural application (such as urban scene understanding or heritage/style analysis) in relation to different scales (from building detail to satellite) and data sources.

3 FINDINGS

3.1 Search results

The compared 88 sources included 29 conference papers, 54 journal papers, 5 online reports, and 1 PhD thesis. Most papers were published in computer science journals (38 articles), followed by journals on remote sensing (13 articles) and architectural and urban studies (8 articles). Table 1 details the thematic distribution of journal disciplines per publication.

3.2 Computer vision in architectural analysis

Visual inspections and analysis are standard practice in building condition surveys and evaluations, with photographs providing a physical record and visual evidence. For example, a visual analysis of façades to determine architectural styles or existing service provisions can be used to establish a building’s age and dwelling type, which might infer typical internal layouts and building maintenance problems. At the same time, building elements such as window or casement types are indicators of thermal performance and are used to establish a building’s EPC rating. The size and location of windows can also provide quantitative information about spatial and environmental qualities including sunlight penetration. Computer vision algorithms can speed and scale up the processing of visual information in cases such as building condition surveys. Generalising, visual data

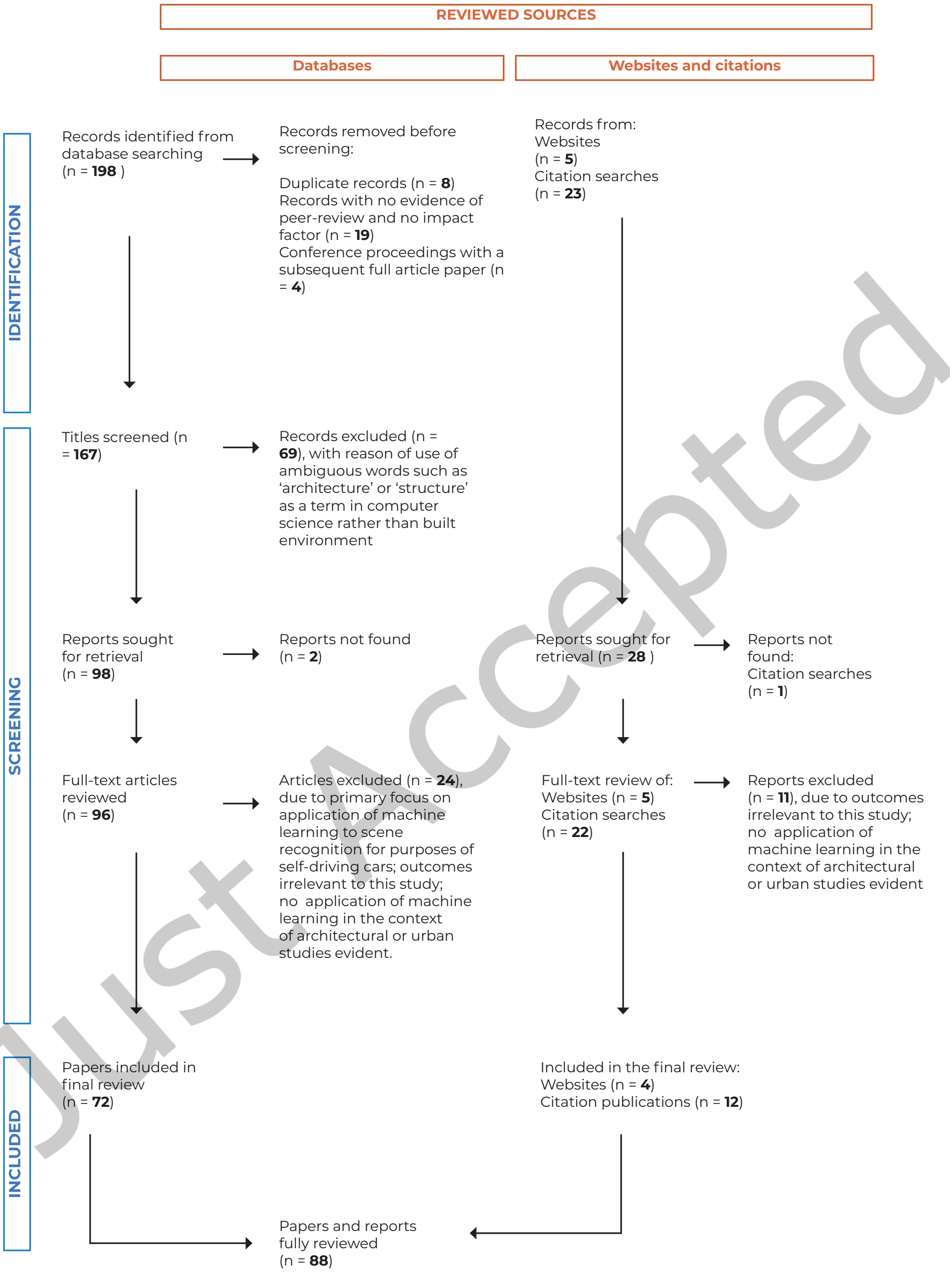

REVIEWED SOURCES
Databases
Websites and citations
IDENTIFICATION
SCREENING
INCLUDED
Records identified from database searching (n = 198)
Records removed before screening:
Duplicate records (n = 8)
Records with no evidence of peer-review and no impact factor (n = 19)
Conference proceedings with a subsequent full article paper (n = 4)
Records from: Websites (n = 5)
Citation searches (n = 23)
Titles screened (n = 167)
Records excluded (n = 69), with reason of use of ambiguous words such as 'architecture' or 'structure' as a term in computer science rather than built environment
Reports sought for retrieval (n = 98)
Reports not found (n = 2)
Reports sought for retrieval (n = 28)
Reports not found: Citation searches (n = 1)
Full-text articles reviewed (n = 96)
Articles excluded (n = 24), due to primary focus on application of machine learning to scene recognition for purposes of self-driving cars; outcomes irrelevant to this study; no application of machine learning in the context of architectural or urban studies evident.
Full-text review of: Websites (n = 5)
Citation searches (n = 22)
Reports excluded (n = 11), due to outcomes irrelevant to this study; no application of machine learning in the context of architectural or urban studies evident
Papers included in final review (n = 72)
Included in the final review: Websites (n = 4)
Citation publications (n = 12)
Papers and reports fully reviewed (n = 88)

Fig. 2. PRISMA flow diagram of source selection for review.

Table 1. Publication categories of the works analysed in this review.

Publication category	Number of publications
Computer science and technology	43
Remote sensing	14
Architecture and urban studies	8
Geography/geoinformation	4
Computer science and design	4
Computer graphics	3
Technology	2
Computer science and architecture	1
Economics	1
VR	1
Environmental Research	1
Natural Sciences	1
PhD Thesis	1
Online sources	4

is currently used in five areas of architectural analysis: building classification, detail classification, qualitative environmental analysis, building condition survey, and building value estimation.

3.2.1 Image classification of building style and typology. Classifying architectural styles and typologies is common in historical and precedent studies in architecture, with the visual analysis of building features part of a process to identify non-visual attributes. Stylistic features might indicate a property's age, region, or construction type. For example, Victorian buildings in England have a distinctive façade design and internal layout. In addition, building uses can be partially indicated by their façade, with computer vision methods utilised to classify architectural styles [62, 71, 108] and typologies [2, 15, 50]. A classification of the whole image is typical for a high-level analysis of architectural features such as overall urban characteristics [41] or style and use classification [60]. Chu and Tsai [16] exploit a graph-mining algorithm to analyse images for repetitive visual patterns that differ between architectural styles. Obeso et al. [76] use a CNN to classify Mexican architectural styles, with visual saliency introduced in the algorithm's network pooling layers to filter relevant features for deeper network layers. Llamas et al. [65] compare the performance of different types of CNNs such as AlexNet or GoogLeNet when trained on pre-labelled images of heritage buildings, and Guo and Li [37] explore improvements to LeNet-5 when applied to architectural style classification tasks.

In another example, the website Classify House A.I. asks users to upload an external image of a house to determine which of the 31 predefined architectural styles can be recognised using computer vision-based analysis [18]. Davies [20] trains an Inception V3 network to recognise Georgian architecture from GSV images of London. Likewise, Alhasoun and González [2] train a CNN to match GSV images to their corresponding US towns by classifying street frontage based on their urban contexts[56]. Deep-learning models are also used by Yoshimura et al. [104] to measure visual similarities between the styles of different architects.

To enhance the architectural benefits of building image classifications, more than one characteristic needs to be considered at the same time, as many exceptions to "ground truth" data can be found across all architectural styles and typologies. For example, dwelling houses might have been converted or changed their use while façades remained the same. Although stylistic and typological features can indicate use and occupancy, they are only one factor, with a more nuanced multi-factor reading needed for reliable estimates.

3.2.2 Building detail detection and classification. Building details, similar to style and typology, can have visual features specific to where a building is located, the local climate, and the period it was built in. For example, high-pitched roofs are historically found in regions affected by heavy snowfall. The detection and classification of building elements involves prior feature extraction and, like style classifications might use indicators such as window designs [87] or face-recognition algorithms (applied to sculpted heads of humans and gargoyles) as a determining feature [86]. A set of stylistic elements extracted from street-view images is used to determine features typical for Paris (or those untypical) in Doersch et al. [25]. In other examples, a bounding-box based object-detection approach separates building details by either extracting whole building façades from an image and then assigning to them a particular style based on their features [100] or by extracting façade details to analyse specific building elements [20, 34, 65, 70, 97, 107]. In other cases, semantic segmentation is applied to detect roof typologies and hedgerows maintenance levels from satellite images [78], to map green and solar roofs [98] or in detail-oriented style analysis where authors train a classifier to distinguish Flemish, Renaissance, Haussmannian, and Neoclassical styles [70].

Of the reviewed papers, almost a third (30 records) discuss semantic segmentation, which is key to extracting elements – either whole façades from their urban context [29, 35, 63, 69] or façade elements such as doors and windows [4, 23, 30, 52, 63, 66, 69, 105, 109]. In research with a focus on semantic segmentation, the extraction of buildings and their elements remains a problem of machine learning techniques and only becomes an architectural question if it is forming part of a larger research process. This includes research on architectural challenges at a scale and complexity difficult to complete using manual methods, such as extracting roof or façade textures to increase the quality of texture patterns in 3D virtual urban models [23], reconstructing urban 3D models [36, 43], or automating building change detection [93].

3.2.3 Qualitative analysis. The exploitation of computer vision in qualitative analysis is still in its infancy, but has noticeably increased in recent years. Most applications assess the quality of streetscapes or establish new links between the aesthetics of an urban environment and statistical data – on education, unemployment, housing, living environment, health, or crime [3, 32, 91]. For example, Streetscore [74] applies Support Vector Regression to predict whether a given streetscape is perceived as safe or unsafe by viewers, and both Dubey et al. [26] and Min et al. [72] study perceptual attributes such as "safe", "lively", "boring", "wealthy", "depressing", and "beautiful" based on GSV images of several US cities. Another study investigates how visual qualities affect how a street is perceived as walkable [103], while Quercia et al. [83] compare the aesthetic qualities of different areas of London. Furthermore, a crowd-labelled dataset of street-view images from Boston and New York is used to create perceptual maps for 21 US cities [58] and the online platform Scenic-Or-Not explores the rating of 200,000 images in relation to perceived qualities of outdoor space, usinga CNN to analyse and extract key features common to positive scores [85]. Ilic et al [46] used a siamese CNN to assess levels of gentrification in Ottawa based on the analysis of properties captured in the GSV images. Similarly, visual preferences are examined through semantic segmentation to understand how individual components such as building façades or greenery relate to perceptions of street space quality [101]. Šćepanović et al. [84] parsed satellite imagery of six Italian cities to predict urban vitality criteria based on the theories of Jane Jacobs. Neighbourhood vitality is also studied by Wang and Vermeulen [96]. Lastly, two recent studies use semantic segmentation and k-means clustering in their urban colour analysis [24, 109]. In the analysed cases, qualitative analysis often requires combining several datasets or applying a multi-stage methodology, or both. This approach to analysis is valuable for making design decisions around a building's form and mass, aesthetics, programme, townscape relationship or user experience, as qualitative assessments are already frequently used in architectural practice.

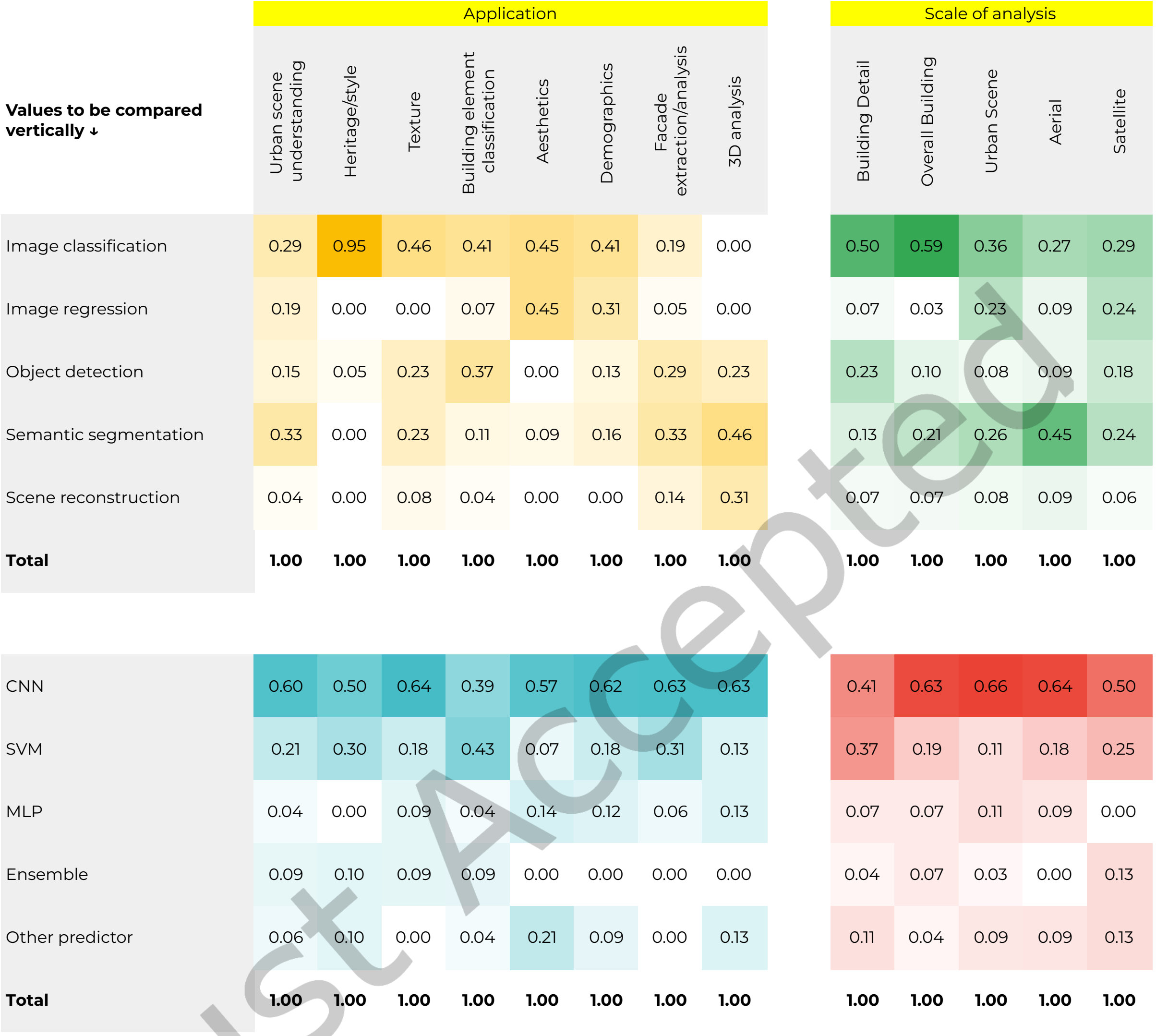
3.2.4 Building condition and value estimation. In the assessment of building conditions and property values image data is an established means for qualitative evaluation of building conditions and a quantitative analysis of building features. Several papers study property price estimation based on a visual assessment. Law et al. [55] use a CNN to automatically extract visual features from GSV images in order to estimate house prices in London, UK. Lindenthal and Johnson [62] combine a traditional hedonic model with architectural style classifications to estimate sales price premia in relation to architectural styles at the building and neighbourhood level, demonstrating that machine learning classifiers are as reliable as human experts in mass appraisals. Wang et al. [95] explore how an aesthetic value might be used to indicate property prices. Similarly, Poursaeed et al. [82] estimate house prices based on visual and textural features, with the dataset including interior and exterior images of buildings that are classified according to levels of perceived luxury. In addition, Muhr et al. [73] use satellite images to automate the assessment of location quality. Computer vision algorithms are further deployed to optimise manual tasks of labelling real estate data. Long short-term memory (LSTM) classification algorithms and fully-connected neural networks (FCNNs) are used in real-estate scene classification to automate the labelling of exterior and interior features that range from room types [12] to countertops [6]. As some of the image data is of insufficient quality, image enhancement processes are also a common feature of this type of research. The use of images in the assessment of building conditions tends to focus on image patch analysis. Examples of this include determining the condition of single-family house based on building elements such as windows or roofs [54]. Zeppelzauer et al. [106] automate a building age estimations using a two-stage approach, first training a CNN to learn the age characteristics at patch level and, second, globally aggregating patch-wise age estimates of an entire building. Another study by Hoang [40] applies SVM to the image analysis of building walls, using cracks in buildings as indicators of fabric deterioration. Visual analysis is a conventional method in building condition and value estimation, with the above studies automating already established processes. The visual building condition survey is used to estimate property values, forecast maintenance costs, or assess a building's state of repair, including dangerous structural deterioration. Consequently, photographs are often used as evidence, for example, in building surveys or tenancy-related inventories.

3.3 Trends and gaps of computer vision in architectural analysis

In the following, the relationship between architectural applications and scales and computer vision tasks and machine learning methods are quantitatively analysed through relative co-occurrence matrices (Fig. 3 and 4). The matrices differ from typical contingency tables as categories are not mutually exclusive. Therefore, a single article may be counted several times and statistical methods such as the chi-squared test cannot be applied to quantify correlations precisely. While a co-occurrence does not necessarily mean that a specific method was used to solve a given problem - only that a method and problem were present at the same time - high contrasts in co-occurrences values suggest a stronger correlation between a particular architectural analysis and computational approach.

Fig. 3 shows the relative occurrences of computational approaches for a given architectural application or scale. Although some applications such as urban scene understanding, demographics, and façade extraction explore a relatively diverse range of computer vision approaches, others including heritage/style and aesthetic analysis have only use specific methods (image classification for the former, classification and regression for the latter). This indicates opportunities to study how other algorithms and methods might perform on these applications, for example, what if style analysis was cast as a regression problem along various perceptual dimensions (classic/modern, rural/urban, etc.) rather than a classification problem? Or, could studies of aesthetics benefit from object detection? While CNNs dominate most architectural applications and scales, consistent with their widespread popularity in other fields, perhaps the more interesting case is when they do not, the classification of building elements (at the building scale), where SVMs are more frequent. One explanation is that such an application typically involves often rectified images of façades with a relatively consistent structure, making the image invariances learned by deep CNNs in exchange for larger amounts of data less useful.

The relative occurrence analysis of architectural applications and scales and computational approaches (Fig. 4) reveals interesting patterns. While computer vision tasks such as image classification, object detection, and

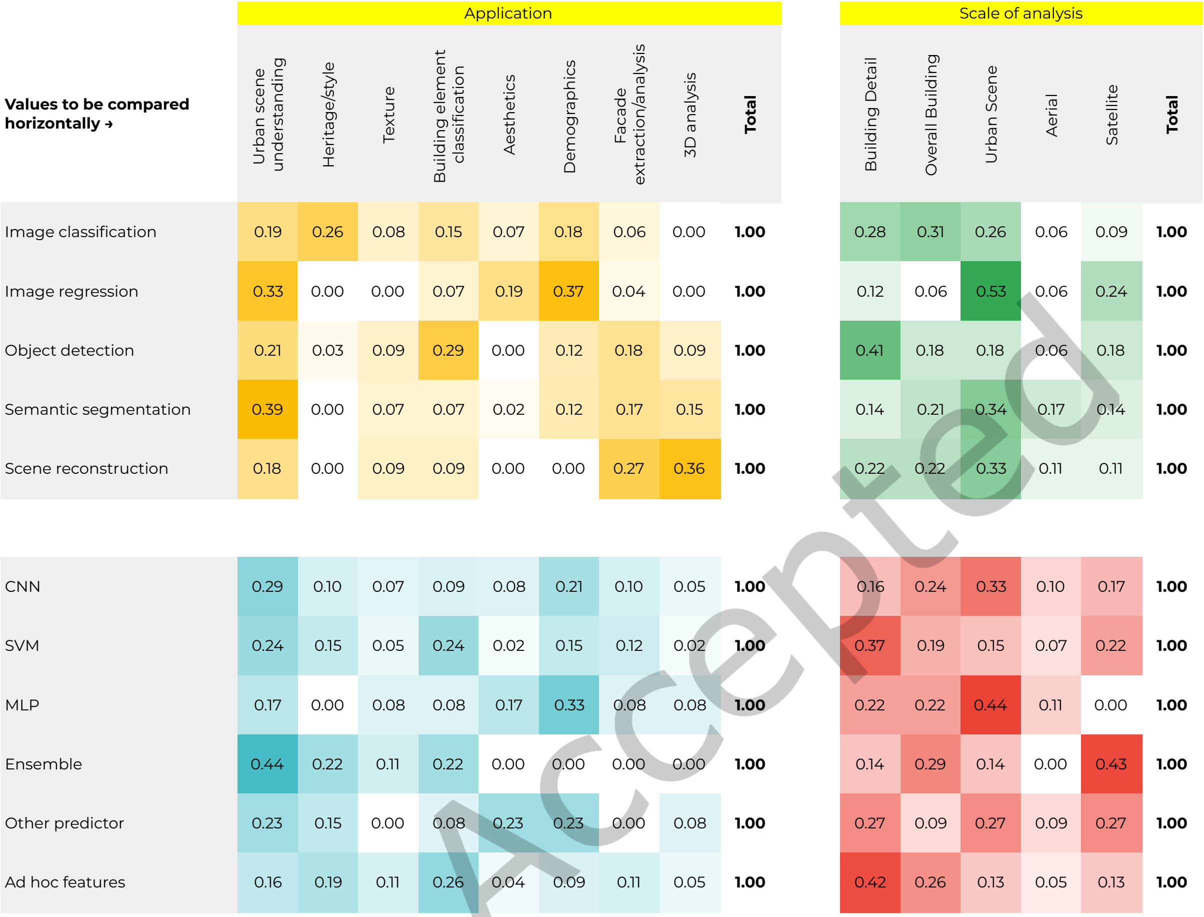


	Application								Scale of analysis				
Values to be compared vertically ↓	Urban scene understanding	Heritage/style	Texture	Building element classification	Aesthetics	Demographics	Facade extraction/analysis	3D analysis	Building Detail	Overall Building	Urban Scene	Aerial	Satellite
Image classification	0.29	0.95	0.46	0.41	0.45	0.41	0.19	0.00	0.50	0.59	0.36	0.27	0.29
Image regression	0.19	0.00	0.00	0.07	0.45	0.31	0.05	0.00	0.07	0.03	0.23	0.09	0.24
Object detection	0.15	0.05	0.23	0.37	0.00	0.13	0.29	0.23	0.23	0.10	0.08	0.09	0.18
Semantic segmentation	0.33	0.00	0.23	0.11	0.09	0.16	0.33	0.46	0.13	0.21	0.26	0.45	0.24
Scene reconstruction	0.04	0.00	0.08	0.04	0.00	0.00	0.14	0.31	0.07	0.07	0.08	0.09	0.06
Total	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
CNN	0.60	0.50	0.64	0.39	0.57	0.62	0.63	0.63	0.41	0.63	0.66	0.64	0.50
SVM	0.21	0.30	0.18	0.43	0.07	0.18	0.31	0.13	0.37	0.19	0.11	0.18	0.25
MLP	0.04	0.00	0.09	0.04	0.14	0.12	0.06	0.13	0.07	0.07	0.11	0.09	0.00
Ensemble	0.09	0.10	0.09	0.09	0.00	0.00	0.00	0.00	0.04	0.07	0.03	0.00	0.13
Other predictor	0.06	0.10	0.00	0.04	0.21	0.09	0.00	0.13	0.11	0.04	0.09	0.09	0.13
Total	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**

Fig. 3. Relative co-occurrences (column-wise). Each value represents the probability of occurrence of a computational method or model given an architectural application or scale.

semantic segmentation are distributed across most architectural applications and scales, others occur in more specific contexts. For instance, image regression is mostly used at the urban scale (i.e., at street or neighbourhood level). This could be due to the difficulty of cross-referencing data sources at smaller scales as current image datasets usually do not identify individual buildings (except for landmarks). Overcoming this challenge might unlock significant potential for the use of image regression in future architectural studies.

Machine learning methods tend to be distributed across all applications and scales, except for ensemble models - although the sample size is too small (n=6) to generalise. Ad hoc features (including SIFT, Haar, HoG, steerable filters, etc.) are a special case in this table, as they are not a predictor but an intermediate representation that is

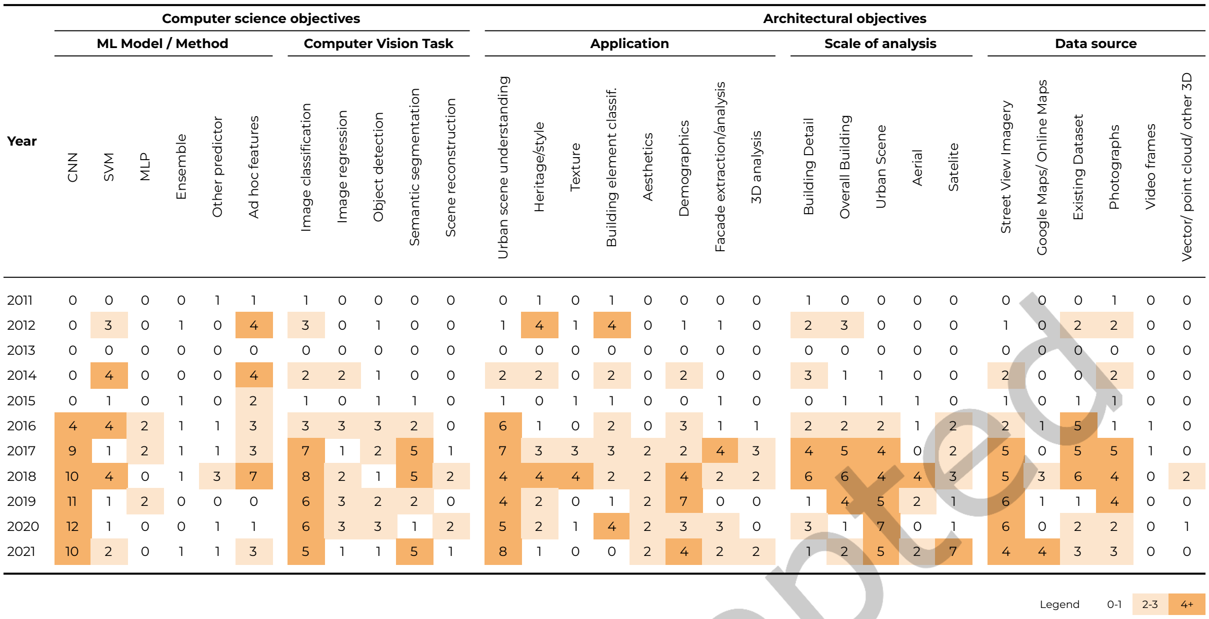


Values to be compared horizontally →	Application									Scale of analysis					
	Urban scene understanding	Heritage/style	Texture	Building element classification	Aesthetics	Demographics	Facade extraction/analysis	3D analysis	**Total**	Building Detail	Overall Building	Urban Scene	Aerial	Satellite	**Total**
Image classification	0.19	0.26	0.08	0.15	0.07	0.18	0.06	0.00	**1.00**	0.28	0.31	0.26	0.06	0.09	**1.00**
Image regression	0.33	0.00	0.00	0.07	0.19	0.37	0.04	0.00	**1.00**	0.12	0.06	0.53	0.06	0.24	**1.00**
Object detection	0.21	0.03	0.09	0.29	0.00	0.12	0.18	0.09	**1.00**	0.41	0.18	0.18	0.06	0.18	**1.00**
Semantic segmentation	0.39	0.00	0.07	0.07	0.02	0.12	0.17	0.15	**1.00**	0.14	0.21	0.34	0.17	0.14	**1.00**
Scene reconstruction	0.18	0.00	0.09	0.09	0.00	0.00	0.27	0.36	**1.00**	0.22	0.22	0.33	0.11	0.11	**1.00**
CNN	0.29	0.10	0.07	0.09	0.08	0.21	0.10	0.05	**1.00**	0.16	0.24	0.33	0.10	0.17	**1.00**
SVM	0.24	0.15	0.05	0.24	0.02	0.15	0.12	0.02	**1.00**	0.37	0.19	0.15	0.07	0.22	**1.00**
MLP	0.17	0.00	0.08	0.08	0.17	0.33	0.08	0.08	**1.00**	0.22	0.22	0.44	0.11	0.00	**1.00**
Ensemble	0.44	0.22	0.11	0.22	0.00	0.00	0.00	0.00	**1.00**	0.14	0.29	0.14	0.00	0.43	**1.00**
Other predictor	0.23	0.15	0.00	0.08	0.23	0.23	0.00	0.08	**1.00**	0.27	0.09	0.27	0.09	0.27	**1.00**
Ad hoc features	0.16	0.19	0.11	0.26	0.04	0.09	0.11	0.05	**1.00**	0.42	0.26	0.13	0.05	0.13	**1.00**

Fig. 4. Relative co-occurrences (row-wise). Each value represents the probability of occurrence of an architectural application or scale given a computational method or model.

input into other models (e.g., SVMs). They were very popular in computer vision before deep learning methods developed and demonstrate the usefulness of older methods. In particular, Figure 4 shows a distribution across all architectural applications and scales, with a stronger co-occurrence at small scale. Again, this can be explained by rectified façade images being more structured, enabling models that use smaller amounts of data to perform relatively well.

To further analyse temporal differences, Fig. 5 breaks down publication numbers for each year across all architectural aspects and computational approaches. The use of computer vision to solve architectural problems such as urban scene understanding, texture, aesthetics, facade, and 3D analysis only seem to become popular with the emergence of deep CNNs. Similarly, studies into heritage, style, and building element analysis - all occurring at larger architectural scales (urban scene and above) - have benefited from earlier computer vision methods (typically SVMs used with ad hoc features for image classification or object detection). This suggests that the ability of deep CNNs to process large amounts of unstructured data has significantly expanded the applications of computer vision to architectural and urban studies.



	Computer science objectives											Architectural objectives																		
	ML Model / Method						Computer Vision Task					Application								Scale of analysis					Data source					
Year	CNN	SVM	MLP	Ensemble	Other predictor	Ad hoc features	Image classification	Image regression	Object detection	Semantic segmentation	Scene reconstruction	Urban scene understanding	Heritage/style	Texture	Building element classif.	Aesthetics	Demographics	Facade extraction/analysis	3D analysis	Building Detail	Overall Building	Urban Scene	Aerial	Satelite	Street View Imagery	Google Maps/ Online Maps	Existing Dataset	Photographs	Video frames	Vector/ point cloud/ other 3D
2011	0	0	0	0	1	1	1	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0
2012	0	3	0	1	0	4	3	0	1	0	0	1	4	1	4	0	1	1	0	2	3	0	0	0	1	0	2	2	0	0
2013	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2014	0	4	0	0	0	4	2	2	1	0	0	2	2	0	2	0	2	0	0	3	1	1	0	0	2	0	0	2	0	0
2015	0	1	0	1	0	2	1	0	1	1	0	1	0	1	1	0	0	1	0	0	1	1	1	0	1	0	1	1	0	0
2016	4	4	2	1	1	3	3	3	3	2	0	6	1	0	2	0	3	1	1	2	2	2	1	2	2	1	5	1	1	0
2017	9	1	2	1	1	3	7	1	2	5	1	7	3	3	3	2	2	4	3	4	5	4	0	2	5	0	5	5	1	0
2018	10	4	0	1	3	7	8	2	1	5	2	4	4	4	2	2	4	2	2	6	6	4	4	3	5	3	6	4	0	2
2019	11	1	2	0	0	0	6	3	2	2	0	4	2	0	1	2	7	0	0	1	4	5	2	1	6	1	1	4	0	0
2020	12	1	0	0	1	1	6	3	3	1	2	5	2	1	4	2	3	3	0	3	1	7	0	1	6	0	2	2	0	1
2021	10	2	0	1	1	3	5	1	1	5	1	8	1	0	0	2	4	2	2	1	2	5	2	7	4	4	3	3	0	0

Legend 0-1 2-3 4+

Fig. 5. Number of publications per year for each computational and architectural aspect.

3.4 Data sources and curation

Data sources can be distinguished in two ways: by their acquisition method and their geographic location. While the former has important implications for the image scale, spatial and temporal resolution, dataset size, and preprocessing and annotations required, the latter directly influences the generalisability of the techniques and findings used.

3.4.1 Acquisition method. Architectural and urban research uses a wide range of image sources and acquisition methods including street-view imagery (36% of studies), photographs scraped, downloaded or taken by the authors (34%), online repositories (32%), online maps (9%), and images extracted from vector and 3D data (4%) or video frames (3%). As noted by [65], there is a lack of image datasets specifically for architectural applications, with researchers often having to create their own: in fact, 84% of the works surveyed in this review did so.

Of the reviewed studies, 25 use photographs both from existing, generic datasets such as ImageNet [32] and building-specific ones such as CMP, eTRIMS or Graz50[1, 42, 64, 66]. This type of dataset can be used for training, but it is also commonly used as an out-of-sample dataset to evaluate the performance of a model. The main advantage of computer vision datasets is that they are readily available for training and evaluation purposes, requiring little preprocessing, as images are already labelled, and for some facade datasets, already rectified. Combining these datasets, however, might require label homogenisation. The main drawback of these existing datasets is that images come with almost no context or metadata such as geographic coordinates or timestamp, making it difficult to combine them with non-visual data sources. Additionally, existing datasets do not currently have a diverse class group needed for a nuanced representation of architectural features and, therefore, are predominantly used in high-level or proof-of-concept studies.

An alternative acquisition approach is to download images from websites following a keyword search, using image repositories such as Flickr [34, 65, 104], Wikimedia [37, 65, 100], Google Image Search [16, 34, 82] or various real estate websites [82, 106]. While the first two sources make finding images easy and permit their use under a CC licence, pictures of landmarks and special interest buildings vastly outnumber those of common buildins. The quality of the metadata is also the most inconsistent of all acquisition methods, because it depends on user annotations, resulting in various misclassification errors. Overall, assembling datasets based on images scraped from image websites requires checks of user-generated tags for accuracy. Real estate websites fare generally better in that regard, but their images are typically protected by copyright [106] and building locations can be purposefully inaccurate for privacy reasons. Some image data is captured by the researchers themselves [86, 94], even including physical synthetic data such as pictures of artificial avian faeces [57]. This creates new opportunities for collaboration when producing custom-made datasets, taking into account both architectural expertise in classifying built environment features and computer vision expertise how to best process the data . While this approach offers optimal control and consistency, the size of the datasets depends on the resources available to the researchers, often resulting in small datasets or highly localised data. While models trained on these images might perform well on the original dataset, generalisability can be poor.

Street-view imagery is the data source of 32 studies, mainly taken from Google Street View but also including alternatives such as Baidu Total View [64, 101, 109]. Images are used at different scales: in their entirety (in streetscape analysis), by extracting individual building façades, or by identifying individual building elements (such as doors and windows). Some use images directly in their panoramic form, others require a rectification step [1, 11]. The pros and cons of street-view imagery for urban research have been discussed at length by Cinnamon and Jahiu in a recent review [17]. Generally, the main advantages of street-view imagery include a rapid data collection at a relatively low cost, dense coverage in some areas, relatively precise geographic coordinates, and the possibility of temporal analyses (although panorama locations are inconsistent over time). The main limitations include occlusions and distortions as well as uneven spatial coverage and frequency of updates.

Satellite and aerial imagery are utilised in 13 and 7 papers respectively, with free-access images sources including the ISPRS 2D Semantic Labeling Contest [5, 49], images collected by the ZY3-01 and JL1-07 satellites [44], and images from the Sentinel satellite programme by the European Space Agency [15, 84]. These sources offer great spatial coverage and precise geographic coordinates, permitting their combination with other data sources. But publicly available remote sensing data often comes only in low resolution. While this is sufficient for the spatial analysis of a whole region, it is not suitable for a analysis of areas smaller than a street block. While the plan view of individual buildings is valuable to gauging general information about size, layout, or materials, the lacking image resolution of public data prevents more detailed studies.

In addition, some studies use of hybrid datasets, combining either different image sources or image data with other data types. Bódis-Szomorú[7] investigates datasets combining street-level and aerial images to automate the updating of 3D urban models. J. Kang et al. [50] supplement remote-sensing data and geographic information with street-level imagery to develop a broader building-use classification based on individual building analysis. Research combining visual with statistical data includes architectural applications of machine learning that previously might not have been possible or evident when using other analytical approaches. For instance, Helber et al. [39] propose a multi-scale machine learning approach to analyse aerial and satellite images in conjunction with socio-economic data to predict property value classes based on image features, Jean et al. [48] pair statistical data of expenditure measurements from the World Bank's Living Standards Measurement Study with Google Static Maps and satellite imagery from the Nighttime Lights Time Series to predict poverty levels, and Su at al. [90] combine high-resolution remote sensing images and statistical data for their urban scene analysis. Hybrid datasets can overcome the limitations of a single data source, but require significant preprocessing to integrate different data.

Generally in the analysed studies greater importance is given to the data processability by computer vision algorithms than the quality and accuracy of architectural representation. Some studies provide very limited information on the image sources [1, 6, 45] and the use of public-domain data without verifying that architectural features are identified correctly is prevalent. Problematic sources include Flickr and Google, as image tags and keywords are often provided by non-experts. None of the research uses available digital image libraries tagged by architectural experts, such as RIBApix (image repository of architectural assets curated by the Royal Institute of British Architects) or the Cities and Buildings Database by the University of Washington. This highlights a need for greater cross-disciplinary collaboration.

3.4.2 Geographic location. Context is essential to understand buildings from an architectural perspective, whether the goal is to assess building style, age, and use, or to combine visual and statistical data. Geographic building location is an especially important information. In the works surveyed, this information is mainly found in custom datasets, but is often provided are at city or country level and only rarely at the building level [110]. No location information is found in 21% of the works. Where the information is given, the location shows a significant concentration of studies in North American, West European, and East Asian countries (Table 2). This bias has implications for the generalisability of the results. For instance, Lotte et al. [66] observe the poor performance of their model trained on mostly European data when applied to a Brazilian context. A limited model transferability might be due to different urban characteristics [13] or architectural styles [52, 69] found in different cities and countries. For example, Nguyen et al. [75] finding that "visible utility wires” are an indicator of physical disorder in the USA are not generalisable, as in countries like Japan the power grid is above ground.

3.5 Reproducibility and comparability

The term “Reproducibility” is defined by the USA’s National Information Standards Organization as the ability to recreate “computational results using the author-created research objects, methods, code, and conditions of analysis” (NISO RP-31-2021). Likewise, the Association for Computing Machinery states: "For computational experiments [...] an independent group can obtain the same result using the author’s own artifacts” (ACM 2020). The reproducibility of the surveyed papers and the comparability of different approaches to the same problem was assessed by checking if the code, custom data, and trained models used by the authors are available. In addition, criteria of reproducibility, specific to machine learning methods, were considered: if the hyperparameters, data splits (training/validation/test), and machine learning software were specified, as well as information such as training and/or inference times and hardware specifications.

The analysed reproducibility criteria are given in Table 3. Each criterion is shown as a percentage of relevant works. For instance, all works rely on code, so there are 88 relevant works for “code” and “hardware”, but 6 do not use machine learning [59, 67, 83, 92, 99, 111], resulting in only 82 counts for “ML software used” and “Training/inference time”. The first three criteria (“code”, “custom data”, and “trained model”) are counted as “fully disclosed” if the artefacts are either available online, upon request, or can be recovered by running a script (for “Custom data”).

The results presented in Table 3 show very low levels of reproducibility: 78% of works are published without any code, 91% without trained models (for those that train their own), and 81% of custom datasets are unavailable (including data that used to be available online but has not been maintained). The results are better for machine learning implementation details: 40% of works provide a full list of hyperparameter values, 80% provide their data splitting strategy (training/validation/test, k-fold cross-validation, or a mixture of both), and 66% disclose the software or framework used. (It is worth noting, however, that the exact version of the software is seldom mentioned, which is problematic when the hyperparameters are said to be left to “default values”.) Training/inference times and hardware specifications are also rarely mentioned.

The comparability of the works was assessed in terms of the diversity of evaluation metrics used for the same type of task. Table 4 shows the number of occurrences of each metric that appeared in more than one paper. Metrics appear to be more variable for image classification and semantic segmentation than for image

Table 2. Occurrence of geographic locations in each custom image dataset. (Hong Kong was considered separately from China because of its specific context in terms of architecture and data access.)

Location	Continent	Occurrences
US	North America	22
UK	Europe	17
France	Europe	12
China	Asia	9
Hong Kong	Asia	6
Canada	North America	5
Germany	Europe	5
Netherlands	Europe	5
Spain	Europe	5
Austria	Europe	4
Italy	Europe	4
Japan	Asia	4
South Korea	Asia	3
Switzerland	Europe	3
Australia	Oceania	2
Belgium	Europe	2
Czech Republic	Europe	2
Denmark	Europe	2
Malawi	Africa	2
Mexico	South America	2
Nigeria	Africa	2
Russia	Europe	2
Rwanda	Africa	2
Singapore	Asia	2
Tanzania	Africa	2
Uganda	Africa	2
Angola	Africa	1
Argentina	South America	1
Benin	Africa	1
Brazil	South America	1
Burkina Faso	Africa	1
Cameroon	Africa	1
Côte d'Ivoire	Africa	1
Democratic Republic of Congo	Africa	1
Ethiopia	Africa	1
Ghana	Africa	1
Greece	Europe	1
Guinea	Africa	1
India	Asia	1
Kenya	Africa	1
Lesotho	Africa	1
Luxembourg	Europe	1
Mali	Africa	1
Mozambique	Africa	1
New Zealand	Oceania	1
Romania	Europe	1
Senegal	Africa	1
Sierra Leone	Africa	1
South Africa	Africa	1
Sweden	Europe	1
Thailand	Asia	1
Togo	Africa	1
Turkey	Europe	1
Ukraine	Europe	1
Vietnam	Asia	1

regression and object detection. This is in part due to the fact that the suitability of a metric depends on the number of classes and their relative proportions. When classes are balanced, accuracy is the simplest metric to compute and compare but many subtle variants exist (such as average accuracy per class or per random split).

Table 3. Disclosure/availability of each reproducibility criterion, as a percentage of works to which the criterion applies.

Reproducibility criterion	Relevant works	Fully disclosed (%)	Partially disclosed (%)	Undisclosed (%)
Code	88	13.6	8.0	78.4
Custom data	74	14.6	4.4	81.1
Trained model	78	5.1	3.8	91.0
Hyperparameters	78	39.7	21.8	38.5
Data split	78	79.5	6.4	14.1
ML software used	82	65.9	N/A	34.1
Training/inference time	82	20.7	6.1	73.2
Hardware	88	36.4	N/A	63.6

Table 4. Occurrence of each evaluation metric.

Computer vision task	Evaluation metric	Occurrences
Image classification	Accuracy	35
	Confusion matrix	16
	F1 score	9
	Precision	8
	Recall	7
	Cohen’s kappa	3
	Error rate	3
	ROC curve / AUC	3
	Kendall’s rank correlation	2
Semantic segmentation	Accuracy	8
	F1 score	6
	Recall	4
	Precision	3
	Confusion matrix	3
Image regression	R^2	7
	MSE	5
Object detection	Average Precision	5

The error rate is the difference between 1 and the accuracy, and its choice remains unusual. The ROC curve (and its associated AUC) is a possible alternative that is more specific to ranked predictions. For classification problems with imbalanced data, a more detailed analysis of the papers reveals a split between those providing the F1 score, precision and recall on one hand, and those using Cohen’s kappa coefficient on the other. Since both F1 and kappa can be used in a binary or multi-class setting (by averaging F1 over classes), this choice appears rather arbitrary. Lastly, confusion matrices are common for multi-class problems, but are difficult to compare across papers because classes are rarely consistent (even for the same task). Overall, while the choice of metric remains task dependent, it is not always consistent across papers solving similar problems, and rarely justified by the authors. Moreover, only 7 papers provide some measure of standard deviation or variance for their metrics, making it more difficult to assess the significance of future improvements.

4 DISCUSSION

4.1 Research trends

There is an extensive application of machine learning in the analysis of architectural features from a computer science perspective. Thereby two significant ways in which recent studies engage with architectural questions and problems can be identified. The first optimises algorithmic methods of image analysis by applying them to image data of architectural or urban environments. This uses both existing and custom-made image recognition models. The primary objective of this type of research is to improve process expediency [77], optimise processing tasks, or enhance accuracy [19, 49]. Once automated, the virtual scene understanding is then deployed in space navigation and virtual visual servoing [97]. The contribution of this kind of research is automating work that is otherwise labour-intensive, thus enabling it to be undertaken faster and at a greater scale and frequency. The shortfall of efficiency-oriented research is that architectural objectives lack sufficient detail and accuracy, and that their focus remains high-level, which can make the results prone to bias.

The second type of research applies existing or custom-made machine learning systems to the analysis of data useful to answering questions arising in the architectural domain, such as a correlation between streetscape aesthetics [83], poverty levels [48] or visual indicators of architectural styles [25, 34, 62, 76, 86, 87, 102]. This includes the qualitative analysis of visual clues and perceived attributes that, for example, might indicate urban gentrification [46] or specific urban environmental conditions [26, 74, 83]. But it also includes the quantitative analysis of buildings and their elements [24], and is sometimes combined with statistical data. Machine learning provides thereby new methods of data analysis with potential for novel understandings of the built environment.

The emerging research that brings together architectural application and computer vision methods is still largely exploratory, and no single dominant interdisciplinary practice has yet emerged. There is thus no standardised evaluation method. There is, consequently, an opportunity to further explore research objectives and processes that are more holistic, collaborative, and interdisciplinary.

4.2 The importance of datasets

Image and hybrid datasets play a key role in the training and testing of computer vision models. Images of buildings are studied from different viewpoints and at different scales, such as street level and elevational or satellite and aerial views, and have a wide range of sources: new photographs, ready-made datasets, scraped data from websites, or collected through custom or public APIs (such as GSV panoramas). Almost none of the reviewed records use the same dataset and data preprocessing and curation needs vary greatly. Moreover, very few of the custom datasets are available online, which might be in part due legal restrictions (copyright law, breach of End User Licence Agreement, etc.).

There are thus challenges around data acquisition and annotation. Crowdsourcing, for instance, can be unreliable. A level of expertise is required to label and class some images, for example, in an architectural dataset [101]. Additionally, data that seems homogeneous might in fact cover different categories. Chu and Tsai [16] classify architecture according to Gothic, Georgian, Korean, and Islamic styles, but while Gothic and Georgian refer to specific art-historical styles and periods, Korean and Islamic are much broader classifications. More generally, the positionality of the annotators is almost never considered, even though it can affect qualitative evaluations such as the definition of “formal” versus “informal” settlements [45] or the collection of “ugly” rooms online [82].

The papers indicate that architectural image datasets are predominantly utilised in computer science studies with a focus on the computational aspects of image analysis and processing and a single database or building scale. From a built environment perspective, however, there is greater value in emphasising urban and architectural analysis, which requires overcoming the current limitation of image datasets that lack context and metadata for an interoperability with other data sources (e.g., statistical surveys).

4.3 Pitfalls of computer vision in architectural analysis

The current lack of reproducibility and comparability in much of the reviewed work indicates that research at the intersection of computer vision and architecture is still very exploratory and emerging. Based on this analysis, three main pitfalls that are particularly prevalent when attempting to tackle architectural issues with a computer vision-based approach can be identified.

The first pitfall is ambiguous training data, which is mentioned in 11 of the works that use image classification. Architecture is rife with typologies and classifications (housing type, use class, style, etc.) that try to capture the diversity of the built environment and involve complex definitions and edge cases. Architectural styles, for instance, are difficult to define precisely [102] as they involve a variety of visual and structural cues, can be combined and transformed and present significant variation across buildings with different use classes. The first step to address this is to ensure that samples are correctly labelled (especially in the case of crowdsourced data). Second, if the error on the training step is unexpectedly high, this might be a case of underfitting: the model does not manage to capture the complex visual relationships that define a class, which might call for a deeper network or a more advanced optimisation algorithm. Third, visual data might simply not be enough to separate the classes unambiguously. In that case, a hybrid dataset should be considered.

The second pitfall is class imbalance, especially common in semantic segmentation and object detection (mentioned in 10 works). It typically occurs in urban scene understanding at aerial or satellite scale, where buildings are small compared to their surroundings or land use classes are unevenly distributed. This problem can be tackled by various strategies, such as minority over-sampling [27], online hard example mining [88], or by modifying the optimisation using, e.g., a focal loss [61].

The third pitfall is view sensitivity, which is a typical problem in street-level analysis. Kim et al. [53] provide a thorough quantitative analysis of the impact of panorama locations on attributes computed from semantic segmentation. However, view sensitivity affects all computer vision tasks due to occlusions, self-occlusions, reflections (for glass buildings), and distortions (for high buildings in narrow streets). Potential approaches to mitigate these problems include using a trained model to detect and remove occluded views, combining visual and non-visual data, or combining inferences from different views of the same building.

5 CONCLUSION: CHALLENGES AND DIRECTIONS OF BUILT ENVIRONMENT RESEARCH

This review demonstrated that studies of the built environment that adopt computer vision methods use two main approaches. The first is to automate classification tasks by mirroring established manual methods of visual analysis, such as the interpretation of architectural or urban elements. Machine learning offers hereby the means to perform quantitative or qualitative analysis at scale. The second approach utilises machine learning as a tool for data processing and analysis to raise novel questions in architectural and urban studies, with the potential for new insights through methodological innovations. Table 5 summarises the identified current research trends and potential future directions.

Future research with built environment applications calls for a greater integration of architectural and computer science methods and aims. For example, automating visual survey methods already in use in architectural practices using computer vision can only make labour-intensive tasks more affordable and data analysis more significant when the visual-spatial criteria and their assessment are well defined and the data used is reliable. Research focused on computer vision problems is unlikely to be relevant to built environment studies unless it is part of a larger research process and questions that integrate disciplinary specificities and analysis. This crossdisciplinary approach has great potential for built environment studies and analysis, making it not only more scalable but also data- and evidence-based, which can ultimately inform design and policy decisions. Shifts in aims and a combination of research and analytical processes across disciplinary boundaries also offer opportunities for methodological innovation and new research outputs with potentially significant real-life impact. The value

Table 5. Current research summary and future directions

Built environment research	
Objectives and applications	Analyse visual-spatial characteristics of building and urban elements for their classification and identification (e.g. styles, typologies, geolocations) Rationalise spatial qualities or quantities by rating visual indicators (e.g. building valuations, building conditions, neighbourhood qualities, walkability) Infer demographic characteristics from visual-spatial analysis (poverty, gentrification). Understand non-physical (virtual) space and generic typologies
Benefits	Automation of labour-intensive work Methodological innovation for novel understanding of built environment based on existing data New insights through transdisciplinary approaches Leveraging existing datasets
Challenges	Generalisability of findings Reproducibility and accuracy of processes Data problems: access, quality, reliability, interoperability, frequency
Future directions	Transdisciplinary approaches to data curation for both qualitative and quantitative visual-spatial analysis Interoperability of visual with other data types to understand and rate spatial relationships or characteristics (e.g. to influence decision-making processes, policy, property value estimation, building condition survey) Integration of multiple feature analysis with different built environment scales (e.g. to predict urban and developmental trends, infer use and occupancy) Democratisation of building-related data
Machine learning objectives	
Test or compare the expediency of classification algorithms Improve the performance and accuracy of image analysis or ML systems (e.g., image segmentation, object or texture detection, object extraction) Test methods to infer information from multiple datasets Identify methods to exploit existing architectural and urban image data	

for architectural research and analysis is numerous: the growing body of research promises benefits for design optimisation, architectural precedent analysis, and policymaking by providing new means of evaluating differences or changes in the built environment.

A key challenge is the availability and processing of data. Differences in data specification, accuracy, interoperability, creation, and access need to be resolved to create comparable and integrated datasets. To achieve this, greater collaboration between different disciplines is needed to ensure that both different expertise and requirements are taken into full consideration. Existing image datasets tend to lack context and metadata beyond GPS coordinates and crowdsourced labels. Resolving data-related problems is especially important when automating the collection of large amounts of data, whether once or at regular intervals. Once this is resolved, analysis that was previously impossible due to highly labour intensive processes can be undertake at ease and repeatedly over time, opening up new forms of analysis and a reliable evidence base currently not available to built environment studies.

Another key challenge is the reproducibility and comparability of computer vision-driven architectural research. While many works so far have been exploratory, future research is likely to seek to confirm current results and improve the state of the art. This requires greater transparency around implementation and data sources. Importantly, it also requires an agreement on assessment metrics meaningful to both computer scientists and architects but also useful for real-world applications related to design decisions or policy.

A particularly promising direction for future research is the integration of visual and non-visual data. Research might consider both building elements and context and combine image sources (like aerial and street views) and other data such as statistical information (e.g. census data, household size, property type information, building age, and socio-economic or environmental statistics). The use of visual indicators in conjunction with existing building-related data, at both a specific point but also over time, has significant potential to create new interdisciplinary approaches, improving on both existing quantitative and qualitative research methods. This can in turn inform design decisions, building safety assessment, maintenance planning, real estate evaluation, and planning or socio-economic policies.

ACKNOWLEDGMENTS

This work was financially supported by the Prosit Philosophiae Foundation.

REFERENCES

[1] Lama Affara, Liangliang Nan, Bernard Ghanem, and Peter Wonka. 2016. Large Scale Asset Extraction for Urban Images. In *Computer Vision – ECCV 2016*. Springer, Cham, 437–452. https://doi.org/10.1007/978-3-319-46487-9_27

[2] Fahad Alhasoun and Marta González. 2019. Urban Street Contexts Classification Using Convolutional Neural Networks and Streets Imagery. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. 1198–1204. https://doi.org/10.1109/ICMLA.2019.00198

[3] Sean M. Arietta, Alexei A. Efros, Ravi Ramamoorthi, and Maneesh Agrawala. 2014. City Forensics: Using Visual Elements to Predict Non-Visual City Attributes. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (Dec. 2014), 2624–2633. https://doi.org/10.1109/TVCG.2014.2346446

[4] Anil Armagan, Martin Hirzer, and Vincent Lepetit. 2017. Semantic segmentation for 3D localization in urban environments. In *2017 Joint Urban Remote Sensing Event (JURSE)*. 1–4. https://doi.org/10.1109/JURSE.2017.7924573

[5] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. 2018. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS Journal of Photogrammetry and Remote Sensing* 140 (June 2018), 20–32. https://doi.org/10.1016/j.isprsjprs.2017.11.011

[6] Jawadul H. Bappy, Joseph R. Barr, Narayanan Srinivasan, and Amit K. Roy-Chowdhury. 2017. Real Estate Image Classification. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE Computer Society, 373–381. https://doi.org/10.1109/WACV.2017.48

[7] András Bódis-Szomorú. 2018. *3D Reconstruction of Urban Scenes from Street-Side and Airborne Imagery.* Doctoral Thesis. ETH Zurich. https://doi.org/10.3929/ethz-b-000286410

[8] Lauren L. Beghini, Alessandro Beghini, Neil Katz, William F. Baker, and Glaucio H. Paulino. 2014. Connecting architecture and engineering through structural topology optimization. *Engineering Structures* 59 (Feb. 2014), 716–726. https://doi.org/10.1016/j.engstruct.2013.10.032

[9] Priyadarshi Bhattacharya and Marina Gavrilova. 2012. A Survey of Landmark Recognition Using the Bag-of-Words Framework. *Intelligent Computer Graphics 2012* (2012), 243–263. https://doi.org/10.1007/978-3-642-31745-3_13

[10] Filip Biljecki and Koichi Ito. 2021. Street view imagery in urban analytics and GIS: A review. *Landscape and Urban Planning* 215 (Nov. 2021), 104217. https://doi.org/10.1016/j.landurbplan.2021.104217

[11] Kirill Bochkarev and Egor Smirnov. 2019. Detecting advertising on building façades with computer vision. *Procedia Computer Science* 156 (Jan. 2019), 338–346. https://doi.org/10.1016/j.procs.2019.08.210

[12] Yang Cao, Shinichi Nunoya, Yusuke Suzuki, Masachika Suzuki, Yoshio Asada, and Hiroki Takahashi. 2019. Classification of real estate images using transfer learning. In *Tenth International Conference on Graphics and Image Processing (ICGIP 2018)*, Vol. 11069. SPIE, 435–440. https://doi.org/10.1117/12.2524417

[13] Bin Chen, Ying Tu, Yimeng Song, David M. Theobald, Tao Zhang, Zhehao Ren, Xuecao Li, Jun Yang, Jie Wang, Xi Wang, Peng Gong, Yuqi Bai, and Bing Xu. 2021. Mapping essential urban land use categories with open big data: Results for five metropolitan areas in the United States of America. *ISPRS Journal of Photogrammetry and Remote Sensing* 178 (Aug. 2021), 203–218. https:

//doi.org/10.1016/j.isprsjprs.2021.06.010

[14] Tao Chen, Kui Wu, Kim-Hui Yap, Zhen Li, and Flora S. Tsai. 2009. A Survey on Mobile Landmark Recognition for Information Retrieval. In *2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware*. 625–630. https://doi.org/10.1109/MDM.2009.107

[15] Wangyang Chen, Abraham Noah Wu, and F. Biljecki. 2021. Classification of Urban Morphology with Deep Learning: Application on Urban Vitality. *Comput. Environ. Urban Syst.* (2021). https://doi.org/10.1016/j.compenvurbsys.2021.101706

[16] Wei-Ta Chu and Ming-Hung Tsai. 2012. Visual pattern discovery for architecture image classification and product image search. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval (ICMR '12)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/2324796.2324831

[17] Jonathan Cinnamon and Lindi Jahiu. 2021. Panoramic Street-Level Imagery in Data-Driven Urban Research: A Comprehensive Global Review of Applications, Techniques, and Practical Considerations. *ISPRS International Journal of Geo-Information* 10, 7 (July 2021), 471. https://doi.org/10.3390/ijgi10070471

[18] Classify [n.d.]. Classify House A.I. http://classifyhouse.com/

[19] Joseph Paul Cohen, Wei Ding, Caitlin Kuhlman, Aijun Chen, and Liping Di. 2016. Rapid building detection using machine learning. *Applied Intelligence* 45, 2 (Sept. 2016), 443–457. https://doi.org/10.1007/s10489-016-0762-6

[20] John Davies. 2019. Street style: machine learning takes to the streets. https://www.nesta.org.uk/report/street-style-machine-learning-takes-streets/

[21] Davide Del Vento and Alessandro Fanfarillo. 2019. Traps, Pitfalls and Misconceptions of Machine Learning applied to Scientific Disciplines. In *Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (learning) (PEARC '19)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3332186.3332209

[22] O. O Demirbaş and H Demirkan. 2003. Focus on architectural design process through learning styles. *Design Studies* 24, 5 (Sept. 2003), 437–456. https://doi.org/10.1016/S0142-694X(03)00013-9

[23] G. Despine and T. Colleu. 2015. Adaptive Texture Synthesis for Large Scale City Modeling. In *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. XL-5-W4. Copernicus GmbH, 155–162. https://doi.org/10.5194/isprsarchives-XL-5-W4-155-2015

[24] Meichen Ding. 2021. Quantitative contrast of urban agglomeration colors based on image clustering algorithm: Case study of the Xia-Zhang-Quan metropolitan area. *Frontiers of Architectural Research* (June 2021). https://doi.org/10.1016/j.foar.2021.05.003

[25] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei A. Efros. 2012. What makes Paris look like Paris? *ACM Transactions on Graphics* 31, 4 (July 2012), 101:1–101:9. https://doi.org/10.1145/2185520.2185597

[26] Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, and César A. Hidalgo. 2016. Deep Learning the City: Quantifying Urban Perception at a Global Scale. In *Computer Vision – ECCV 2016*. Springer, Cham, 196–212. https://doi.org/10.1007/978-3-319-46448-0_12

[27] Zhou Fang, Jiaxin Qi, Tianren Yang, Li Wan, and Ying Jin. 2020. "Reading" cities with computer vision: a new multi-spatial scale urban fabric dataset and a novel convolutional neural network solution for urban fabric classification tasks. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '20)*. Association for Computing Machinery, New York, NY, USA, 507–517. https://doi.org/10.1145/3397536.3422240

[28] Tian Feng, Feiyi Fan, and Tomasz Bednarz. 2021. A review of computer graphics approaches to urban modeling from a machine learning perspective. *Frontiers of Information Technology & Electronic Engineering* 22, 7 (July 2021), 915–925. https://doi.org/10.1631/FITEE.2000141

[29] Antoine Fond, Marie-Odile Berger, and Gilles Simon. 2017. Facade Proposals for Urban Augmented Reality. In *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 32–41. https://doi.org/10.1109/ISMAR.2017.20

[30] Antoine Fond, Marie-Odile Berger, and Gilles Simon. 2021. Model-image registration of a building's facade based on dense semantic segmentation. *Computer Vision and Image Understanding* 206 (May 2021), 103185. https://doi.org/10.1016/j.cviu.2021.103185

[31] Sourav Garg, Tobias Fischer, and Michael Milford. 2021. Where is your place, Visual Place Recognition? *IJCAI* (2021). https://doi.org/10.24963/ijcai.2021/603

[32] Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. 2017. Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. *Proceedings of the National Academy of Sciences* 114, 50 (Dec. 2017), 13108–13113. https://doi.org/10.1073/pnas.1700035114

[33] Jorge Gil, José Nuno Beirão, Nuno Montenegro, and Jose M. Pinto Duarte. 2012. On the discovery of urban typologies: Data mining the many dimensions of urban form. *Urban Morphology* 16, 1 (March 2012), 27–40. http://www.scopus.com/inward/record.url?scp=84858401417&partnerID=8YFLogxK

[34] Abhinav Goel, Mayank Juneja, and C. V. Jawahar. 2012. Are buildings only instances? Exploration in architectural style categories. In *Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP '12)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/2425333.2425334

[35] Fang-Ying Gong, Zhao-Cheng Zeng, Fan Zhang, Xiaojiang Li, Edward Ng, and Leslie K. Norford. 2018. Mapping sky, tree, and building view factors of street canyons in a high-density urban environment. *Building and Environment* 134 (April 2018), 155–167.

https://doi.org/10.1016/j.buildenv.2018.02.042

[36] Shengxi Gui and Rongjun Qin. 2021. Automated LoD-2 model reconstruction from very-high-resolution satellite-derived digital surface model and orthophoto. *ISPRS Journal of Photogrammetry and Remote Sensing* 181 (Nov. 2021), 1–19. https://doi.org/10.1016/j.isprsjprs.2021.08.025

[37] Kun Guo and Ning Li. 2017. Research on classification of architectural style image based on convolution neural network. In *2017 IEEE 3rd Information Technology and Mechatronics Engineering Conference (ITOEC)*. 1062–1066. https://doi.org/10.1109/ITOEC.2017.8122517

[38] Nan He and Guanghao Li. 2021. Urban neighbourhood environment assessment based on street view image processing: A review of research trends. *Environmental Challenges* 4 (Aug. 2021), 100090. https://doi.org/10.1016/j.envc.2021.100090

[39] Patrick Helber, Benjamin Bischke, Qiushi Guo, Jörn Hees, and Andreas Dengel. 2019. Multi-Scale Machine Learning for the Classification of Building Property Values. In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*. 4873–4876. https://doi.org/10.1109/IGARSS.2019.8900257

[40] Nhat-Duc Hoang. 2018. Image Processing-Based Recognition of Wall Defects Using Machine Learning Approaches and Steerable Filters. *Computational Intelligence and Neuroscience* 2018 (Nov. 2018), e7913952. https://doi.org/10.1155/2018/7913952

[41] Chuan-Bo Hu, Fan Zhang, Fang-Ying Gong, Carlo Ratti, and Xin Li. 2020. Classification and mapping of urban canyon geometry using Google Street View images and deep multitask learning. *Building and Environment* 167 (Jan. 2020), 106424. https://doi.org/10.1016/j.buildenv.2019.106424

[42] H. Hu, L. Wang, M. Zhang, Y. Ding, and Q. Zhu. 2020. Fast and Regularized Reconstruction of Building Façades from Street-View Images Using Binary Integer Programming. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. V-2-2020. Copernicus GmbH, 365–371. https://doi.org/10.5194/isprs-annals-V-2-2020-365-2020

[43] Zhihua Hu, Yaolin Hou, Pengjie Tao, and Jie Shan. 2021. IMGTR: Image-triangle based multi-view 3D reconstruction for urban scenes. *ISPRS Journal of Photogrammetry and Remote Sensing* 181 (Nov. 2021), 191–204. https://doi.org/10.1016/j.isprsjprs.2021.09.009

[44] Xin Huang, Junjing Yang, Jiayi Li, and Dawei Wen. 2021. Urban functional zone mapping by integrating high spatial resolution nighttime light and daytime multi-view imagery. *ISPRS Journal of Photogrammetry and Remote Sensing* 175 (May 2021), 403–415. https://doi.org/10.1016/j.isprsjprs.2021.03.019

[45] Mohamed R Ibrahim, James Haworth, and Tao Cheng. 2021. URBAN-i: From urban scenes to mapping slums, transport modes, and pedestrians in cities using deep learning and computer vision. *Environment and Planning B: Urban Analytics and City Science* 48, 1 (Jan. 2021), 76–93. https://doi.org/10.1177/2399808319846517

[46] Lazar Ilic, M. Sawada, and Amaury Zarzelli. 2019. Deep mapping gentrification in a large Canadian city using deep learning and Google Street View. *PLOS ONE* 14, 3 (March 2019), e0212814. https://doi.org/10.1371/journal.pone.0212814

[47] Farzad Jalaei, Milad Zoghi, and Afshin Khoshand. 2021. Life cycle environmental impact assessment to manage and optimize construction waste using Building Information Modeling (BIM). *International Journal of Construction Management* 21, 8 (Aug. 2021), 784–801. https://doi.org/10.1080/15623599.2019.1583850

[48] Neal Jean, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. 2016. Combining satellite imagery and machine learning to predict poverty. *Science* 353, 6301 (Aug. 2016), 790–794. https://doi.org/10.1126/science.aaf7894

[49] Michael Kampffmeyer, Arnt-Børre Salberg, and Robert Jenssen. 2016. Semantic Segmentation of Small Objects and Modeling of Uncertainty in Urban Remote Sensing Images Using Deep Convolutional Neural Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 680–688. https://doi.org/10.1109/CVPRW.2016.90

[50] Jian Kang, Marco Körner, Yuanyuan Wang, Hannes Taubenböck, and Xiao Xiang Zhu. 2018. Building instance classification using street view images. *ISPRS Journal of Photogrammetry and Remote Sensing* 145 (Nov. 2018), 44–59. https://doi.org/10.1016/j.isprsjprs.2018.02.006

[51] Yuhao Kang, Fan Zhang, Song Gao, Hui Lin, and Yu Liu. 2020. A review of urban physical environment sensing using street view imagery in public health studies. *Annals of GIS* 26, 3 (July 2020), 261–275. https://doi.org/10.1080/19475683.2020.1791954

[52] Tom Kelly, John Femiani, Peter Wonka, and Niloy J. Mitra. 2017. BigSUR: large-scale structured urban reconstruction. *ACM Transactions on Graphics* 36, 6 (Nov. 2017), 204:1–204:16. https://doi.org/10.1145/3130800.3130823

[53] Jae Hong Kim, Sugie Lee, John R. Hipp, and Donghwan Ki. 2021. Decoding urban landscapes: Google street view and measurement sensitivity. *Computers, Environment and Urban Systems* 88 (July 2021), 101626. https://doi.org/10.1016/j.compenvurbsys.2021.101626

[54] David Koch, Miroslav Despotovic, Muntaha Sakeena, Mario Döller, and Matthias Zeppelzauer. 2018. Visual Estimation of Building Condition with Patch-level ConvNets. In *Proceedings of the 2018 ACM Workshop on Multimedia for Real Estate Tech (RETech'18)*. Association for Computing Machinery, New York, NY, USA, 12–17. https://doi.org/10.1145/3210499.3210526

[55] Stephen Law, Brooks Paige, and Chris Russell. 2019. Take a Look Around: Using Street View and Satellite Images to Estimate House Prices. *ACM Transactions on Intelligent Systems and Technology* 10 (Sept. 2019), 1–19. https://doi.org/10.1145/3342240

[56] Stephen Law, Chanuki Illushka Seresinhe, Yao Shen, and Mario Gutierrez-Roig. 2020. Street-Frontage-Net: urban image classification using deep convolutional neural networks. *International Journal of Geographical Information Science* 34, 4 (April 2020), 681–707. https://doi.org/10.1080/13658816.2018.1555832

[57] Jiseok Lee, Jooyoung Hong, Garam Park, Hwa Soo Kim, Sungon Lee, and Taewon Seo. 2020. Contaminated Facade Identification Using Convolutional Neural Network and Image Processing. *IEEE Access* 8 (2020), 180010–180021. https://doi.org/10.1109/ACCESS.2020.

3027839

[58] Lezhi Li, J. Tompkin, P. Michalatos, and H. Pfister. 2017. Hierarchical Visual Feature Analysis for City Street View Datasets. https://www.semanticscholar.org/paper/Hierarchical-Visual-Feature-Analysis-for-City-View-Li-Tompkin/24d0092a3fc2a414b73fd453c5cdd229fc0cb174

[59] Xiaojiang Li and Carlo Ratti. 2019. Mapping the spatio-temporal distribution of solar radiation within street canyons of Boston using Google Street View panoramas and building height model. *Landscape and Urban Planning* 191 (2019), 103387. https://doi.org/10.1016/j.landurbplan.2018.07.011

[60] Xiaojiang Li, Chuanrong Zhang, and Weidong Li. 2017. Building block level urban land-use information retrieval based on Google Street View images. *GIScience & Remote Sensing* 54, 6 (Nov. 2017), 819–835. https://doi.org/10.1080/15481603.2017.1338389

[61] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal Loss for Dense Object Detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 2999–3007. https://doi.org/10.1109/ICCV.2017.324

[62] Thies Lindenthal and Erik B. Johnson. 2021. Machine Learning, Architectural Styles and Property Values. *The Journal of Real Estate Finance and Economics* (July 2021). https://doi.org/10.1007/s11146-021-09845-1

[63] Hantang Liu, Yinghao Xu, Jialiang Zhang, Jianke Zhu, Yang Li, and Steven C. H. Hoi. 2020. DeepFacade: A Deep Learning Approach to Facade Parsing With Symmetric Loss. *IEEE Transactions on Multimedia* 22, 12 (Dec. 2020), 3153–3165. https://doi.org/10.1109/TMM.2020.2971431

[64] Lun Liu, Elisabete A. Silva, Chunyang Wu, and Hui Wang. 2017. A machine learning-based method for the large-scale evaluation of the qualities of the urban environment. *Computers, Environment and Urban Systems* 65 (2017), 113–125. https://doi.org/10.1016/j.compenvurbsys.2017.06.003

[65] Jose Llamas, Pedro M. Lerones, Roberto Medina, Eduardo Zalama, and Jaime Gómez-García-Bermejo. 2017. Classification of Architectural Heritage Images Using Deep Learning Techniques. *Applied Sciences* 7, 10 (Oct. 2017), 992. https://doi.org/10.3390/app7100992

[66] Rodolfo Georjute Lotte, Norbert Haala, Mateusz Karpina, Luiz Eduardo Oliveira e Cruz de Aragão, and Yosio Edemir Shimabukuro. 2018. 3D Façade Labeling over Complex Scenarios: A Case Study Using Convolutional Neural Network and Structure-From-Motion. *Remote Sensing* 10, 9 (Sept. 2018), 1435. https://doi.org/10.3390/rs10091435

[67] Yi Lu. 2019. Using Google Street View to investigate the association between street greenery and physical activity. *Landscape and Urban Planning* 191 (2019), 103435. https://doi.org/10.1016/j.landurbplan.2018.08.029

[68] Lei Ma, Yu Liu, Xueliang Zhang, Yuanxin Ye, Gaofei Yin, and Brian Alan Johnson. 2019. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing* 152 (June 2019), 166–177. https://doi.org/10.1016/j.isprsjprs.2019.04.015

[69] Markus Mathias, Anđelo Martinović, and Luc Van Gool. 2016. ATLAS: A Three-Layered Approach to Facade Parsing. *International Journal of Computer Vision* 118, 1 (May 2016), 22–48. https://doi.org/10.1007/s11263-015-0868-z

[70] M. Mathias, A. Martinovic, J. Weissenberg, S. Haegler, and L. Van Gool. 2012. Automatic Architectural Style Recognition. In *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. XXXVIII-5-W16. Copernicus GmbH, 171–176. https://doi.org/10.5194/isprsarchives-XXXVIII-5-W16-171-2011

[71] Marina Adriana Mercioni and Stefan Holban. 2019. A study on Hierarchical Clustering and the Distance metrics for Identifying Architectural Styles. In *2019 International Conference on ENERGY and ENVIRONMENT (CIEM)*. 49–53. https://doi.org/10.1109/CIEM46456.2019.8937662

[72] Weiqing Min, Shuhuan Mei, Linhu Liu, Yi Wang, and Shuqiang Jiang. 2020. Multi-Task Deep Relative Attribute Learning for Visual Urban Perception. *IEEE Transactions on Image Processing* 29 (2020), 657–669. https://doi.org/10.1109/TIP.2019.2932502

[73] Valentin Muhr, Miroslav Despotovic, David Koch, Mario Döller, and Matthias Zeppelzauer. 2017. Towards Automated Real Estate Assessment from Satellite Images with CNNs14-2. In *10th Forum Media Technology*. 14–22.

[74] Nikhil Naik, Jade Philipoom, Ramesh Raskar, and César Hidalgo. 2014. Streetscore – Predicting the Perceived Safety of One Million Streetscapes. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE Computer Society, 793–799. https://doi.org/10.1109/CVPRW.2014.121

[75] Quynh C. Nguyen, Yuru Huang, Abhinav Kumar, Haoshu Duan, Jessica M. Keralis, Pallavi Dwivedi, Hsien-Wen Meng, Kimberly D. Brunisholz, Jonathan Jay, Mehran Javanmardi, and Tolga Tasdizen. 2020. Using 164 Million Google Street View Images to Derive Built Environment Predictors of COVID-19 Cases. *International Journal of Environmental Research and Public Health* 17, 17 (Sept. 2020), E6359. https://doi.org/10.3390/ijerph17176359

[76] Abraham Montoya Obeso, Jenny Benois-Pineau, Alejandro Álvaro Ramirez Acosta, and Mireya Saraí García Vázquez. 2016. Architectural style classification of Mexican historical buildings using deep convolutional neural networks and sparse features. *Journal of Electronic Imaging* 26, 1 (Dec. 2016), 011016. https://doi.org/10.1117/1.JEI.26.1.011016

[77] Abraham Montoya Obeso, Jenny Benois-Pineau, Mireya Saraí García Vázquez, and Alejandro A. Ramírez Acosta. 2018. Introduction of Explicit Visual Saliency in Training of Deep CNNs: Application to Architectural Styles Classification. In *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*. 1–5. https://doi.org/10.1109/CBMI.2018.8516465

[78] Andrew Orlowski. 2017. UK's map maker Ordnance Survey plays with robo roof detector. https://www.theregister.com/2017/12/14/ordnance_survey_ml_experiments/

[79] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *International Journal of Surgery* 88 (April 2021), 105906. https://doi.org/10.1016/j.ijsu.2021.105906

[80] Matthew J. Page, David Moher, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and Joanne E. McKenzie. 2021. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ* 372 (March 2021), n160. https://doi.org/10.1136/bmj.n160

[81] Brady Peters and Terri Peters. 2018. *Computing the Environment: Digital Design Tools for Simulation and Visualisation of Sustainable Architecture*. John Wiley & Sons.

[82] Omid Poursaeed, Tomáš Matera, and Serge Belongie. 2018. Vision-based real estate price estimation. *Machine Vision and Applications* 29, 4 (May 2018), 667–676. https://doi.org/10.1007/s00138-018-0922-2

[83] Daniele Quercia, Neil O'Hare, and Henriette Cramer. 2014. Aesthetic capital: What makes London look beautiful, quiet, and happy? 945–955. https://doi.org/10.1145/2531602.2531613

[84] Sanja Šćepanović, Sagar Joglekar, Stephen Law, and Daniele Quercia. 2021. Jane Jacobs in the Sky: Predicting Urban Vitality with Open Satellite Data. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 48:1–48:25. https://doi.org/10.1145/3449257

[85] Chanuki Illushka Seresinhe, Tobias Preis, and Helen Susannah Moat. 2017. Using deep learning to quantify the beauty of outdoor places. *Royal Society Open Science* 4, 7 (2017), 170170. https://doi.org/10.1098/rsos.170170

[86] Gayane Shalunts, Martin Cerman, and Daniel Albertini. 2017. Detection of sculpted faces on building facades. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. 677–685. https://doi.org/10.1109/APSIPA.2017.8282119

[87] Gayane Shalunts, Yll Haxhimusa, and Robert Sablatnig. 2011. Architectural Style Classification of Building Facade Windows. In *Advances in Visual Computing*. Springer, Berlin, Heidelberg, 280–289. https://doi.org/10.1007/978-3-642-24031-7_28

[88] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. 2016. Training Region-based Object Detectors with Online Hard Example Mining. *arXiv:1604.03540 [cs]* (April 2016). http://arxiv.org/abs/1604.03540

[89] Nizam Onur Sönmez. 2018. A review of the use of examples for automating architectural design tasks. *Computer-Aided Design* 96 (March 2018), 13–30. https://doi.org/10.1016/j.cad.2017.10.005

[90] Yu Su, Yanfei Zhong, Qiqi Zhu, and Ji Zhao. 2021. Urban scene understanding based on semantic and socioeconomic features: From high-resolution remote sensing imagery to multi-source geographic datasets. *ISPRS Journal of Photogrammetry and Remote Sensing* 179 (Sept. 2021), 50–65. https://doi.org/10.1016/j.isprsjprs.2021.07.003

[91] Esra Suel, John W. Polak, James E. Bennett, and Majid Ezzati. 2019. Measuring social, environmental and health inequalities using deep learning and street imagery. *Scientific Reports* 9, 1 (April 2019), 6229. https://doi.org/10.1038/s41598-019-42036-w

[92] Jakub T. Szcześniak, Yu Qian Ang, Samuel Letellier-Duchesne, and Christoph F. Reinhart. 2022. A method for using street view imagery to auto-extract window-to-wall ratios and its relevance for urban-level daylighting and energy simulations. *Building and Environment* 207 (2022), 108108. https://doi.org/10.1016/j.buildenv.2021.108108

[93] Aparna Taneja, Luca Ballan, and Marc Pollefeys. 2015. Geometric Change Detection in Urban Environments Using Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 11 (Nov. 2015), 2193–2206. https://doi.org/10.1109/TPAMI.2015.2404834

[94] Deepank Verma, Arnab Jana, and Krithi Ramamritham. 2019. Machine-based understanding of manually collected visual and auditory datasets for urban perception studies. *Landscape and Urban Planning* 190 (Oct. 2019), 103604. https://doi.org/10.1016/j.landurbplan.2019.103604

[95] Feng Wang, Yang Zou, Haoyu Zhang, and Haodong Shi. 2019. House Price Prediction Approach based on Deep Learning and ARIMA Model. In *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*. IEEE Computer Society, 303–307. https://doi.org/10.1109/ICCSNT47585.2019.8962443

[96] Mingshu Wang and Floris Vermeulen. 2021. Life between buildings from a street view image: What do big data analytics reveal about neighbourhood organisational vitality? *Urban Studies* 58, 15 (Nov. 2021), 3118–3139. https://doi.org/10.1177/0042098020957198

[97] Fan Wei, Yuan Li, and Lior Shamir. 2019. Computer Analysis of Architecture Using Automatic Image Understanding. *Journal of Data Mining & Digital Humanities* 2018 (Jan. 2019), 4683. https://doi.org/10.46298/jdmdh.4683

[98] Abraham Noah Wu and Filip Biljecki. 2021. Roofpedia: Automatic mapping of green and solar roofs for an open roofscape registry and evaluation of urban sustainability. *Landscape and Urban Planning* 214 (Oct. 2021), 104167. https://doi.org/10.1016/j.landurbplan.2021.104167

[99] Jing Xiao, Markus Gerke, and George Vosselman. 2012. Building extraction from oblique airborne imagery based on robust façade detection. *ISPRS Journal of Photogrammetry and Remote Sensing* 68 (2012), 56–68. https://doi.org/10.1016/j.isprsjprs.2011.12.006

[100] Zhe Xu, Dacheng Tao, Ya Zhang, Junjie Wu, and Ah Chung Tsoi. 2014. Architectural Style Classification Using Multinomial Latent Logistic Regression. In *Computer Vision – ECCV 2014*. Springer, Cham, 600–615. https://doi.org/10.1007/978-3-319-10590-1_39

[101] Yu Ye, Wei Zeng, Qiaomu Shen, Xiaohu Zhang, and Yi Lu. 2019. The visual quality of streets: A human-centred continuous measurement based on machine learning algorithms and street view images. *Environment and Planning B: Urban Analytics and City Science* 46, 8 (Oct. 2019), 1439–1457. https://doi.org/10.1177/2399808319828734

[102] Yun Kyu Yi, Yahan Zhang, and Junyoung Myung. 2020. House style recognition using deep convolutional neural network. *Automation in Construction* 118 (Oct. 2020), 103307. https://doi.org/10.1016/j.autcon.2020.103307

[103] Li Yin and Zhenxin Wang. 2016. Measuring visual enclosure for street walkability: Using machine learning algorithms and Google Street View imagery. *Applied Geography* 76 (Nov. 2016), 147–153. https://doi.org/10.1016/j.apgeog.2016.09.024

[104] Yuji Yoshimura, Bill Cai, Zhoutong Wang, and Carlo Ratti. 2019. Deep Learning Architect: Classification for Architectural Design Through the Eye of Artificial Intelligence. In *Computational Urban Planning and Management for Smart Cities*, Stan Geertman, Qingming Zhan, Andrew Allan, and Christopher Pettit (Eds.). Springer International Publishing, Cham, 249–265. https://doi.org/10.1007/978-3-030-19424-6_14

[105] Zhiliang Zeng, Mengyang Wu, Wei Zeng, and Chi-Wing Fu. 2020. Deep Recognition of Vanishing-Point-Constrained Building Planes in Urban Street Views. *IEEE Transactions on Image Processing* 29 (2020), 5912–5923. https://doi.org/10.1109/TIP.2020.2986894

[106] Matthias Zeppelzauer, Miroslav Despotovic, Muntaha Sakeena, David Koch, and Mario Döller. 2018. Automatic Prediction of Building Age from Photographs. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval (ICMR '18)*. Association for Computing Machinery, New York, NY, USA, 126–134. https://doi.org/10.1145/3206025.3206060

[107] Luming Zhang, Mingli Song, Xiao Liu, Li Sun, Chun Chen, and Jiajun Bu. 2014. Recognizing architecture styles by hierarchical sparse coding of blocklets. *Information Sciences* 254 (Jan. 2014), 141–154. https://doi.org/10.1016/j.ins.2013.08.020

[108] Peipei Zhao, Qiguang Miao, Jianfeng Song, Yutao Qi, Ruyi Liu, and Daohui Ge. 2018. Architectural Style Classification Based on Feature Extraction Module. *IEEE Access* 6 (2018), 52598–52606. https://doi.org/10.1109/ACCESS.2018.2869976

[109] Teng Zhong, Cheng Ye, Zian Wang, Guoan Tang, Wei Zhang, and Yu Ye. 2021. City-Scale Mapping of Urban Façade Color Using Street-View Imagery. *Remote Sensing* 13, 8 (Jan. 2021), 1591. https://doi.org/10.3390/rs13081591

[110] Peihao Zhu, Wamiq Reyaz Para, Anna Fruehstueck, John Femiani, and Peter Wonka. 2020. Large Scale Architectural Asset Extraction from Panoramic Imagery. *IEEE Transactions on Visualization and Computer Graphics* 28, 2 (2020), 1301–1316. https://doi.org/10.1109/TVCG.2020.3010694

[111] Qing Zhu, Mier Zhang, Han Hu, and Feng Wang. 2020. Interactive Correction of a Distorted Street-View Panorama for Efficient 3-D Façade Modeling. *IEEE Geoscience and Remote Sensing Letters* 17, 12 (2020), 2125–2129. https://doi.org/10.1109/LGRS.2019.2962696