CLIP and the City: Addressing the Artificial Encoding of Cities in Multimodal Foundation Deep Learning Models

Dario Negueruela del Castillo, Iacopo Neri
Digital Visual Studies, University of Zurich
dario.negueruelaldelcastillo@uzh.ch
iacopo.neri@uzh.ch

Abstract: In this project, we propose and explore a computational pipeline to examine urban cultural landscapes through the lens of artificial intelligence, and for questioning modes of embedding culture in machine learning models. By employing machine learning models that extract features and textual properties from images, we aim to uncover the connections between a city's history, architecture, and urban development. The city of Rome serves as a significant case study for this research.
To achieve this objective, we feed 360° panoramic images into large vision-language models (e.g. OpenCLIP), to question how mainstream culture is expressed in these models. In this machine-triggered urban experiment, we investigate overlaps between history and machinic interpretation and whether relevant temporal correlations can be captured through generic street images only. Finally, by spatially analysing the captured data, we identify clusters and discontinuities in the urban layout aiming at visually depicting the interplay of forces behind its development.
As in a forensic exercise, the paper seeks to uncover the complex social and historical dynamics of urban environments, exploiting only contemporary images of their settings and a generic embedding of culture. It explores potential cultural biases embedded in machine learning models by comparing Rome - culturally relevant for the western world - with other cities around the world; leveraging innovative computational pipelines and globally covering datasets to provide a novel research line for urban studies.

Keywords: AI, Urban Studies, Multimodality, Computational Mapping, CLIP

1 Introduction

Cities are crucial man-made environments. Sites of cultural and economic innovation, they are also dynamic and living iconic repositories of memory and identity. Moreover, cities populate our dreams of futures to either cherish or avoid.
On the other hand, cities have long exceeded human control or comprehension, operating according to their own complex rhythms and logics. As Henri Lefebvre observed, the urban constitutes an autonomous force that cannot be fully grasped through traditional categories like the political, economic or social (Lefebvre, 2003). Rather, as Read argues, cities emerge through infrastructures that gather and distribute heterogeneous elements, translating global virtualities into local actualities (Read, 2013).

Figure 1 Process image for multimodal urban studies

In this light, we propose a computational pipeline for investigating the cultural landscape of cities through machine vision (Figure 1). By feeding panoramic images into multimodal models based on CLIP (Contrastive Language-Image Pre-training), we explore alternative applications for such technology in the field of urban studies, simultaneously questioning the extension to which images alone can be representative of urbanity and revealing how these models embed cultural assumptions or biases. Questions of cultural identity, architectural history, urban economics are addressed among others to test the performance of the model in disparate fields and, consequently, to question its universal application. Beyond a mere theoretical exercise, this becomes

relevant in light of the operational image, as discussed by Farocki (2004). Serving as the knowledge base for an increasing number of Machine Learning models, CLIP architectures possess the potential to directly influence real-world contexts through its visual preferences, finally deserving extensive consideration. Crucially, this experiment views the city not as a bounded human artefact, but as a provisional point of articulation in flexible networks, continuously remade over time. The models extract visual features and textual properties, but miss the unseen continuities that pattern individual lives, as Read describes (2013). By spatially analysing the captured data, we seek to identify clusters and discontinuities expressing the interplay of forces behind the development of the city under study.

Our approach goes beyond surface appearances to probe the virtual centralities that precede and make visible place possible. It conceives the urban as a found world, appropriated and reworked, rather than a fixed social construct (Massey, 2005). As such, it promises novel perspectives on entrenched assumptions regarding culture, history and technology, opening and sketching a new direction for the study of cities.

2 Background

Cities are where we spend most of our lives and therefore our visual environment - the imagerial background if you wish - to our lives. So, what can we learn from looking at its appearance beyond questions of architectural style and urban form? Can the image of cities be a synthetic testimony to its underlying social and spatial processes and logics? This type of inquiry only emerged strongly in the second half of the XX century when key urban scholars witnessed and addressed rapidly changing urban landscapes. Their pioneering contributions still inspire contemporary studies, providing an evolving and adaptable source of knowledge and tactics, which deserves a concise summary.

2.1 Seeing the City

In postwar central Europe, members of the Internationale Situationiste, influenced by avant-garde movements like Dadaism, developed creative practices for critically engaging with urban space, including the *dérive* (drifting subversively through the city) and practices of detournement to subvert the spectacle of consumerist landscapes. Deeply compelling, their daring and visionary methods and interventions - despite limited in scale and continuity (Plant, 1992) - remain a source of inspiration for addressing how the atmosphere and the mental geography or psychogeography impact our behaviour and spur us on to the continued exploration of alternative approaches for studying the city.

It is however since Kevin Lynch's pioneering work in "The Image of the City" (1960) that urban scholars have increasingly turned to visual and sensory methods for studying the complex life of cities in a more scientific manner. Lynch examined how residents perceive urban environments through *mental maps* oriented around paths, edges, districts, nodes and landmarks. While providing insights into urban navigation and design, however, Lynch's approach has been critiqued for focusing excessively on physical forms over social processes behind the cityscape (Tonkiss, 2013).

Jane Jacobs expanded sensory approaches in "The Death and Life of Great American Cities" (1961), advocating first-hand observation of neighbourhoods as key resources for urban studies. She valued the messy vitality of the street and its complex intermingling of uses and users, not without receiving criticism. Her pioneering contribution, seen as romanticising urban diversity, has been questioned for ignoring exclusionary dimensions beneath surface-level mixing (Brenner et al., 2012).

Influenced by Marxist philosophy, Henri Lefebvre conceived of space as produced through recursive interactions between social relations, ideologies, and physical form. In "The Production of Space" (1991), Lefebvre highlighted how spatial practices and built environments shape everyday urban life. Lefebvre's social theorization of space inspired extensive scholarship specifically on key concepts of spatial justice and right to the city, though some argue it overemphasised abstract domination over human agency (Soja, 1996). A little later, in "The Practice of Everyday Life", Michel de Certeau (1984) examined how residents tactically reappropriate urban spaces through subversive pedestrian practices like walking, eluding visual capture by institutional strategies of control. However, critics argue de Certeau romanticises individualised resistance in the absence of larger structural change (Buchanan, 2000). Yet afterwards, Sharon Zukin vividly illuminated connections between culture and the urban political economy in "The Cultures of Cities" (1995), studying how public debates over architecture, gentrification and aesthetics shape urban identities. Her positions have been accused of overstating causal claims regarding the cultural impact of deindustrialization (Ley, 1996).

While varying in approach, these perspectives collectively demonstrate how cities take form and meaning not just through material structures but also through contested fields of vision, imagination, spectacle, strategy, and representation. The urban landscape condenses social histories and relations, requiring contextualised visual

analysis to unpack. However, limitations around structure versus agency, romanticization, scale, causation, and intersectionality ought to be addressed.

2.2 The Computational Eye

In parallel, a much more quantitative approach was paving the way for groundbreaking technologies and city scale disaster. Seminal works like the one of William H. Whyte, particularly in his 1980's book "The Social Life of Small Urban Spaces", laid the foundation for empirical urban studies through meticulous first-eye observations, and consolidated the relevance of data gathering to comprehend and effectively manage urban areas (Whyte, 1980). Despite providing an effective and replicable method to analyse human-space interaction, Whyte's approach required spending countless hours in public spaces, meticulously recording and annotating, and soon was no longer viable at scale (Batty, 2013).

As time marched on, the birth of data science and novel developments in computing technologies brought an ever increasing attention to data in all fields of application, and the urban realm was not an exception (Batty, 2013; Kitchin, 2014). Occasionally culminating in tantalising concepts like the one of *smart cities*, the promise of data in optimising complex urban environments has been driving economic interests in city planning for decades (Greenfield, 2013). Around the turn of the 21st century, cities all over the world seemed impelled to start implementing IoT technologies, from sensors to monitor air quality and traffic to CCTV cameras for security reasons, fostering a global digital-oriented transformation and literally becoming city-scale information systems (Batty, 2013). While proven insufficient in inherently leading to a conclusive understanding of urban life due to heavy reliance on corporate interests, and issues of data ethics and privacy, these experiments offered fertile ground for data analytics to thrive (Greenfield, 2013; Kitchin, 2014).

In the midst of growing scepticism towards the smart city concept, a glimmer of optimism emerged in the form of computer vision models. These state-of-the-art technologies held the potential to harness advanced algorithms and machine learning for facilitating visual observations (Zhu et al., 2017). In contrast to human observers, computer vision models operated without the constraints of time, rendering them a formidable instrument for scrutinising public spaces (Naik et al., 2014). Nevertheless, despite their considerable potential, these models had their own set of limitations. State of the Art vision models (e.g. YOLO[1]) excelled in tasks like object detection and enumeration (Redmon et al., 2016), yet they struggled to encapsulate the nuanced social, cultural, and contextual aspects that shape the tapestry of urban life (Gebru et al., 2021). Finally, urban studies stand at an intriguing crossroads, as qualitative and quantitative approaches have each unveiled relevant structuring facets of cities, yet both retain certain limitations (Brenner, 2013; Jacobs, 1961). However, the emergence of multimodal AI models signals a profound shift in analysing intricate urban complexity (Saxena et al., 2022). These models integrate the strengths of existing methods while mitigating singular limitations, as they offer a new promise for computationally deciphering cities through a fusion of data-driven precision and nuanced understanding (Young et al., 2022).

§The computational eye, with its quantitative computational urban analysis has significantly expanded the understanding of physical urban dynamics, from flows to formations (Batty, 2013). However, purely data-driven approaches often overlook nuanced cultural, social, political aspects that are constitutive and crucial to urban life (Jacobs, 1961). Conversely, influential qualitative perspectives from scholars like Jane Jacobs and Henri Lefebvre have timely illuminated experiential and sociological urban complexity, but they nevertheless encounter challenges of scale and discourse-dependent subjectivity (Brenner, 2013).

In the next chapter, we delve into this transformation as we introduce the advent of the multimodal CLIP neural network and conduct an inquiry into its capacities for understanding urban environments. The model's ability to process extensive datasets, coupled with its potential capacity to interpret some of the multifaceted nature of cities, positions it as a crucial candidate in the quest to unravel urban research.

2.3 The Emergence of Multimodal Foundation Models

From 2018, following the introduction of BERT[2], AI has been undergoing a paradigm shift with the rise of the so-called foundation models (Bommasani et al., 2021), trained on large, broad datasets, which enables them to be adapted for a wide variety of downstream tasks. The term *foundation model* underscores how these models have become a fundamental building block in the modular stack of many AI systems.

The development of these models is enabled by a method called *transfer learning*, which allows models pre-trained on broad data to transfer knowledge to specialised tasks. It is the scale of data and model size is what gives foundation models their powerful capabilities. The concept originated in natural language processing (NLP)

[1] See https://github.com/ultralytics/ultralytics

[2] See https://github.com/google-research/bert

with influential models like BERT and ChatGPT demonstrating the potential and stimulating the development of similar models such as SimCLR (Chen et al., 2020a), MoCo (He et al., 2020), BEiT (Bao et al., 2022), and MAE (He et al., 2022a). In recent years, the foundation model paradigm has gained traction, expanding beyond NLP to computer vision and other domains. As models are trained on ever-growing data at ever-larger scales, foundation models have become a centralised pillar of machine learning while still requiring further specialisation for more task-specific applications.

Multimodality, with its fusion of data-driven computational power and nuanced qualitative understanding, offers a tantalising prospect as "a new kind of telescope and microscope" for understanding multifaceted phenomena (Manovich, 2017, p. 332). Multimodal models aim at synthesising the dualism quantitative/qualitative, leveraging vast data analysis while inferring nuanced human contexts (Saxena et al., 2022)[3]. Their agnostic nature circumvents preconceived biases that can limit human analyses (DeVries et al., 2022). Specifically, the Contrastive Language-Image Pre-training (CLIP)[4] model developed by OpenAI in the early 2021, demonstrates unique potential to computationally decipher urban environments through integrating language and visual data (Radford et al., 2021). Moving beyond siloed inputs, CLIP's dual encoders unlock new capacities to process extensive heterogeneous urban data while interpreting complex cultural contexts (Gabriel et al., 2022).

3 Multimodal Urban Studies

To address the performance of CLIP models for urban studies, we set up an experiment potentially replicable for any context around the world. Leveraging globally covering street image datasets like Google Street View, we work under the premise that observational data captured in public spaces in the form of panoramic images can act as a proxy for deeper urban dynamics. By comparing several locations within a city against a series of labels or terms via OpenCLIP[5], we aim at a distant reading of the public space mediated only by its culturally biassed training dataset[6], therefore agnostic enough to be applicable for different research questions. Compared to other studies involving CLIP models, the innovation of the proposed methodology lies in the consequent spatialization of the inferred evaluations, which are read at a city scale and provide a novel perspective that possibly highlights recognizable concentrations in the form of clusters or patterns. As a consequence, the city is considered not as a constellation of punctual events but an intertwined patchwork of styles and layers.

In this regard, many are the aspects that are being questioned through this experiment - currently serving more as a replicable manifesto than a severe methodology, - from established knowledge on cities, to the machine learning model involved and their urban imaginaries. Embracing a reflective practice, in the best tradition of urban studies, the aim of this exercise is not to find an optimised answer, but, instead, to trigger better questions.

3.1 Encoding Process and Urban Imaginaries

The encoding process within CLIP is akin to many deep learning models that learn to recognize patterns, relationships, and correlations within data through a process known as feature extraction (LeCun et al., 2015). As the model processes the myriad images of urban landscapes and their accompanying textual descriptors, it engages in a complex multi-dimensional mapping. The data is mapped onto a high-dimensional latent space where proximity denotes similarity in features or characteristics (Goodfellow et al., 2016). The dimensions of this latent space could potentially capture a multitude of facets of urban life such as architectural styles, spatial configurations, or socio-cultural interactions. Over time, this encoding process distils the essence of urban imaginaries into a multi-dimensional construct, creating a rich, albeit abstracted, representation of urban spaces (Zeiler & Fergus, 2014).

The latent space within CLIP serves as a digital mirror, reflecting the collective urban imaginaries encapsulated in its training data. Mircea Eliade, posited that cities are physical manifestations of a society's cosmology (Eliade, 1957); and in a similar vein, the latent space of CLIP reflects the digital cosmology of urban spaces as captured in the training data. This, however, is not a mere replica but a filtered and distilled representation (Bengio et al., 2013). The diversity of architectural styles, the rhythm of urban layouts, and the dynamism of socio-cultural interactions find a machinic abstract representation within the latent space. The range of urban narratives, from

[3] Even if they are, in our opinion, currently far from the structured and articulated complexity required for in-depth qualitative analysis.

[4] See https://github.com/openai/CLIP

[5] Open source model based on OpanAi's CLIP. See https://laion.ai/blog/large-openCLIP/

[6] Which consists of unfiltered, highly varied, and highly noisy data, publicly available on the internet. See https://openai.com/research/CLIP. This set arguably originated in the Common Crawl. For reference the latest crawl dating June 2023, is 390 TiB. see https://commoncrawl.org/overview

the sprawling territorial urbanity to the dense metropolises, is encoded in a manner that can be navigated and explored algorithmically (Zheng et al., 2019). The abstracted representations within the latent space can be perceived as a form of computational phenomenology, where the essence of urban experiences is captured and encoded (McCormack et al., 2019)

3.2 The computational pipeline

Technically, for the assessment of CLIP models' multifaceted and debated urban comprehension, we present in the following a Python-based computational pipeline. Organised around two main sections: setup and assessment, the pipeline automates the collection of observational data for the urban context under study, and consequently runs visual assessments at street scale against given prompts to finally plot the results into a map (Figure 2).

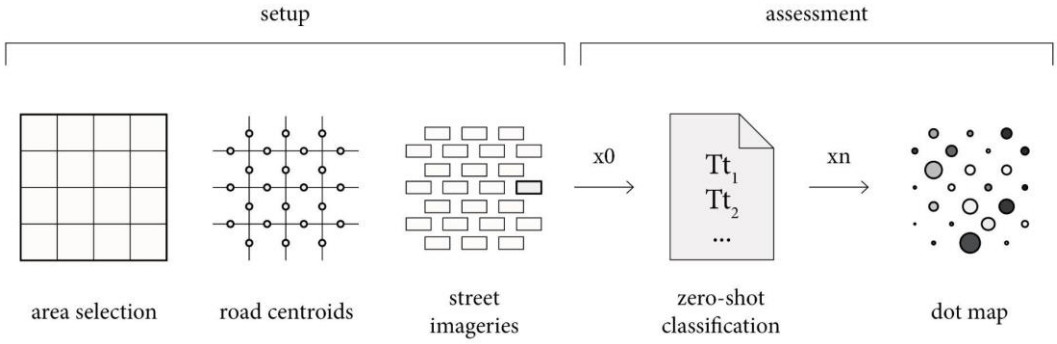

Figure 2 Diagram of the computational pipeline for multimodal urban studies

Point of the departure for any study, the setup phase is responsible for the creation of the imagerial dataset that becomes the primary synthetic representation of the analysed city. Aiming at maximising replicability, it permits comparison studies between contexts towards both inter-cities assessments as well as inter-cultural evaluation of OpenCLIP. In this regard, two main data sources are exploited to avoid issues of availability of local data:

- OpenStreetMap (OSM)[7]: to geocode a location via the Nomatim tool, and to retrieve a vectorial base layer of the road network
- Google Street View (GSV)[8]: to access street level panoramic images for the coordinates extracted via OSM

As roads become the preferred entry points from where to look at the city, the pipeline takes advantage of the OSMnx library[9] that not only dynamically accesses OSM data but also provides a flexible toolkit to clean and prepare street networks. An initial representation of the urban context is obtained via the *graph_from_place()* function subfiltered on the *primary, secondary, tertiary, residential,* and *pedestrian* categories. Consequently, using the GeoPandas[10] library for geospatial data manipulation, the network is reduced into a collection of representative locations lying at the centroid of each segment and waiting to be filled with panoramic images via the Google Street View Static API[11].

Once the setup is over and the imagerial database is created, the visual assessment with OpenCLIP can begin. The sole input here being a set of labels, properly processed into textual prompts, as the methodology leverages the application of CLIP models for zero-shot classification. Technically, this implies the ability of the model to classify unseen images into unseen categories by simply generalising from its prior learning, having embedded both the image and the prompt. By evaluating cosine similarity for each image against all prompts, the most relevant pairs are possibly inferred without need for fine-tuning. For the case study introduced in this paper, extensive label sets for "cities" or "architects" are iteratively introduced in prompts like "a panoramic street image of {city}" or "a panoramic photo of a street designed by {architect}" to investigate different scenarios.

Lastly once each image has been processed, the most pertinent labels alongside the model's similarity scores are geospatially mapped into their original coordinates. Resulting in a dot map where colours and radii are

[7] See https://www.openstreetmap.org

[8] See https://www.google.com/streetview/

[9] See https://osmnx.readthedocs.io/en/stable/

[10] See https://geopandas.org/en/stable/

[11] See https://developers.google.com/maps/documentation/streetview?hl=it

respectively representative of labels and scores, the final output provides an additional lens to read the model's evaluations that goes beyond the correctness of the single assessment, in favour of the collective pattern emerging from the underlying spatial distribution.

3.3 Rome seen through CLIP

How would then a multimodal foundation model, with its limitations as a general learner trained on the mainstream popular culture of the internet, see and characterise Rome? Avoiding the futility of asking for very precise or fine-grained questions, which also risk being irrelevant in the face of the large and multifaceted urban phenomenon, we choose to approach the inquiry through a set of questions that involve a more or less clear thematic polarity, as to explicitly reflect valorisation of urban appearance by the model. In this regard, we depart from a simple scoring approach to later evaluate classification tasks with standardised taxonomies. Finally, we conclude with two provocative experiments to support everlasting questions on the identity and culture of cities. Before we begin with the assessments, however, it is important to contextualise Rome within the framework mentioned above. Spanning over an area of 1.284 km^2, Rome offers an incredibly rich and varied context within its municipal boundaries where centuries of developments have endured. For the purposes of this paper, nonetheless, a case study is created in relation to its historical core area only, more precisely the Rioni on the eastern side of the Tevere river. Here, an initial street network is accessed in the form of 7.143 street segments and separated from 4.454 segments of *footways* and *trackways*, generally unequipped with GSV data. This is further reduced into 4.157 locations where proximity between centroids is less than 20 metres, to obtain a satisfying balance in the results between resolution and relevance. Thus, the panoramas collected represent a wide array of urban features, spanning from iconic architectures to less trafficked alleys, and collectively provide an image of a *real* Rome that goes beyond the solely tourist attractions.

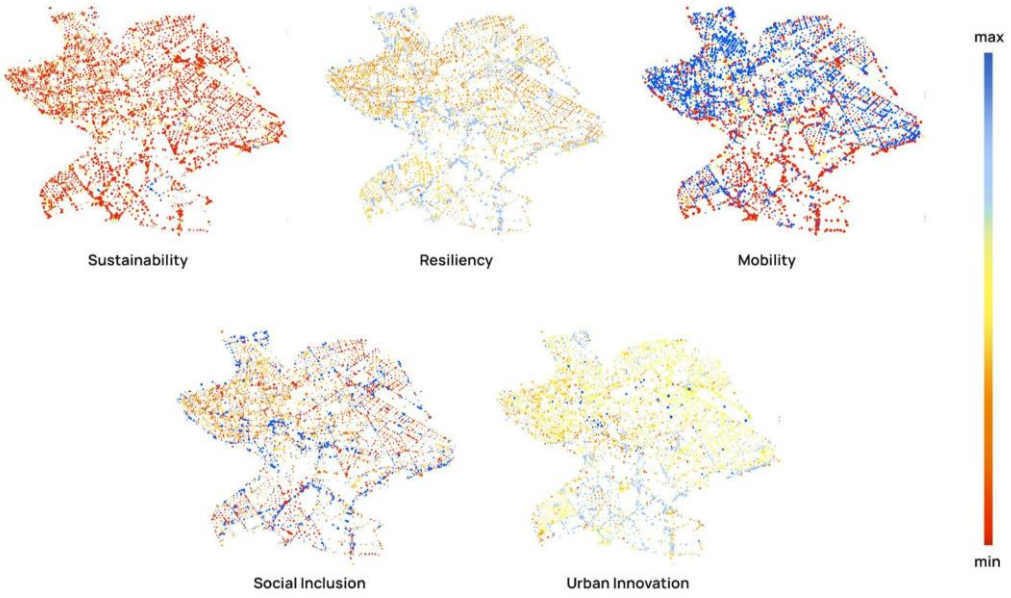

Figure 3 Dot maps resulting from the scoring evaluation for the topics of Sustainability, Resiliency, Mobility, Social Inclusion, Urban Innovation. Colours and radii are respectively showcasing the most pertinent performance labels and the similarity scores with which the model assessed them.

We begin our enquiry following a simple scoring logic aiming at assessing key performance indicators for sustainable urban development. Taking inspiration from international reports on innovative practices for city planning[12], five label sets ranging in five degrees from low to high performance are produced around the topics of: sustainability, resiliency, mobility, social inclusion and urban innovation. The results (Figure 3) reveal a remarkable diversity in the area under study, both within each indicator as well as in their cross-reading, and preliminarily prove the applicability of the approach. This becomes particularly relevant when considering the resulting spatial distribution, which clearly escapes pure randomness. In this regard and acknowledging decades of visual urban studies, Rome is undoubtedly best understood through the experiment as a multi-layered

[12] See the thematic tracks of the European Commission's Intelligent Cities Challenge (ICC), https://www.intelligentcitieschallenge.eu/

tapestry of urban features, rather than an arbitrary constellation of isolated events. On the other hand, it is important to ponder on the wide spectrum that the adopted terminologies embrace as they function like guidelines for policymakers to initiate decision-making processes, and not like specific, granular analysis. Thus, further experiments are warranted to confirm the validity of each chapter in the stance of supporting serious policy frameworks; more specifically in regard to the adopted labels and the process of prompting, which solely derived from the authors' understanding of the subject.

Precisely the latter issue becomes a matter of investigation for a following experiment where the label set is extracted from a standardised taxonomy. Exploiting the CORINE Urban Atlas Land Cover/Land Use[13] categories, we evaluate the capability of the model to infer the land use patterns, bridging ground-level observations with established terminologies used in environmental studies. Answering the call for a consistent and comprehensive dataset to map land across nations, the CORINE (Coordination of Information on the Environment) program provides a successful example of land monitoring service at the European level that has been producing large-scale and comparable land cover, biotope, and air quality maps since the early 1990s. Among the highest resolution outputs that have been developed, the Urban Atlas is a metropolitan scale dataset covering 788 cities[14] at the finest scale of 0.25 ha. Adopting the same taxonomy but exploiting street-level images instead of satellite ones, the study aims to provide an alternative lens to discuss these datasets toward the goal of ubiquitous replicability. Additionally, and in line with traditional land use maps, the results are aggregated into a coarser grid that leaves no void within the area under study and highlights the most recurrent labels among close locations. Once again, the main focus of the study is shifted from the single evaluation to its collective distribution as two clear polarities emerge between the southern and the northern sections of the area, respectively under the *other roads and associated land* and *continuous urban fabric* labels. Hinting at a diffused difference in the current facades, urban furnitures, and usages of the area, these results are tangentially aligned with the administrative subdivisions of Rome - the Rioni - and might be representative of deeper historical conditions. Amidst the remaining cells escaping the above-mentioned two labels, the area of Roma Termini's train station gets interestingly classified as a port area, triggering a promising quantitative point of departure for future studies on non-places (Augé, 1992).

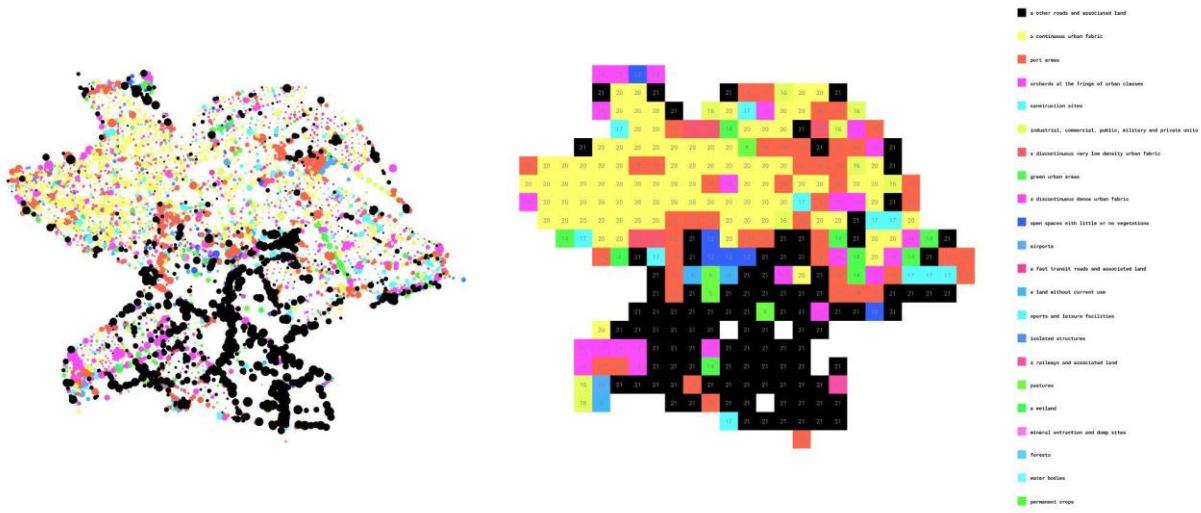

Figure 4 Results of the classification task using the Corine Land Cover taxonomy. The image illustrates the process of resampling with the values being aggregated into a coarser resolution. Similarity scores are not considered for the resampling, which only evaluates the occurrences of each label

[13] See https://doi.org/10.2909/fb4dffa1-6ceb-4cc0-8372-1ed354c285e6

[14] Precisely, the dataset refers to Functional Urban Areas as the urban area resulting from aggregation of a city and its commuting zone. Functional urban areas therefore consist of a densely inhabited city and a less densely populated commuting zone whose labour market is highly integrated with the city. See https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Functional_urban_area#:~:text=Short%20definition%3A%20a%20functional%20urban,city%20(OECD%2C%202012)

Contrary to the application-focus we have discussed until now, a more sophisticated question arises in the following section where architects and cities are evaluated against the synthetic representation of Rome to shed a light into OpenCLIP's cultural understanding. Serving as profound cultural proxies, we argue that architects and cities are ideal labels for the discussed methodology, as they encapsulate intricate relationships with the urban, simultaneously bearing an intimate connection to the aesthetics and visual characteristics depicted in the panoramic images.

Thus, we decided to query the model for confidence in associating the panoramas with a selection of 211 Italian architects[15]. Spanning centuries and from different regions of origin, the results show Mattia De Rossi (1637-1695, Rome), Aloisio the New (XVI century, probably Venetian or Lombardian but worked mostly in Crimea and Russia), and Florestano di Fausto (1890-1965, Rome, epitome of Italian colonial architecture), achieving the three highest scores overall (Figure 5). Agreeably, this assigns a baroque architect to majorly represent the core of Rome, yet surprisingly not the most notable one. Moreover, a puzzling behaviour emerges at a deeper look, as Mattia De Rossi - despite being the most recurrent choice - is also overall reported with very low similarity scores. We therefore believe the label Mattia De Rossi to be an entry point for the exploration of the Rome that lies in the backstage of its monumental and iconic architectures, as he seems to encapsulate the undefined complexity of fuzzy contexts, seamlessly unrelatable to other well-known architects[16].

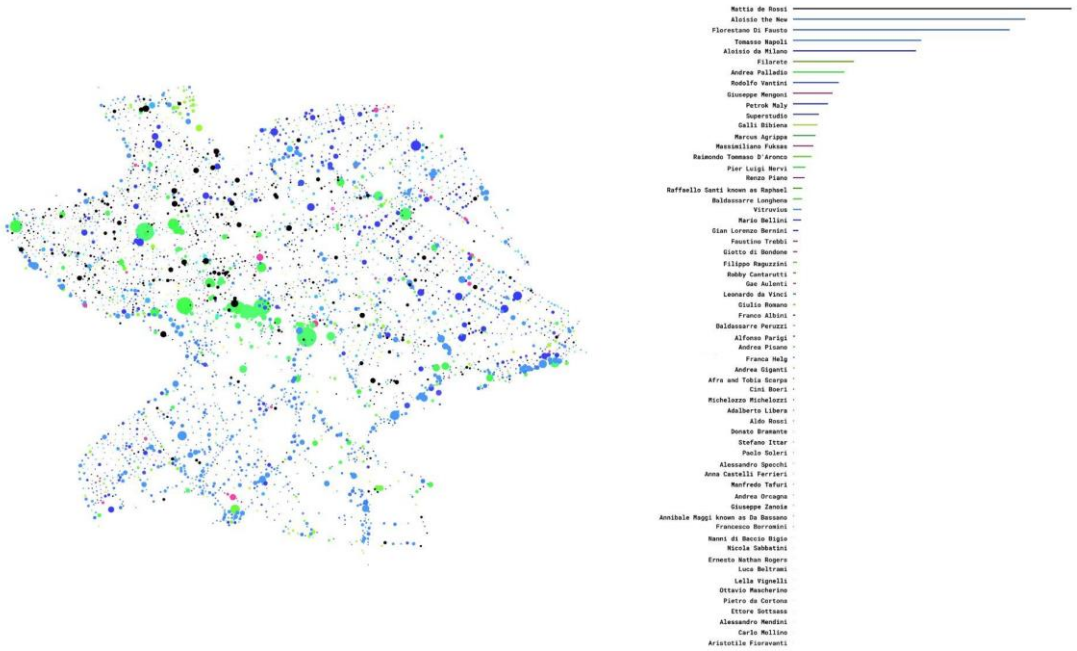

Figure 5 Dot map showcasing the distribution of the most identified Italian architects. Mattia de Rossi is the only entry reported in black while the radii are in function of the similarity score given by the model. On the right, a bar chart shows the values distribution.

Along the same line, a multiscalar study is carried out to position Rome in the global worldview of OpenCLIP. Beginning with a comparison of Rome to other Italian provinces, we gradually enlarged our focus to include European capitals and, ultimately, world capitals (Figure 6). The results show an indicative performance as the model struggles to correctly identify Rome in relation to the Italian provinces (achieving 18th place), to suddenly reach a consensus at the European scale (achieving 1st place) that is only maintained at the global scale if aggregated with the Vatican City label (their sum achieving 1st place). Interestingly, this offers a parallel with human perception as discerning Rome within the Italian context can be said to be a much more challenging task than compared with further distant contexts and permits cross-scale studies to precisely pinpoint views of Rome that can be arguably recognized as its mainstream image.

[15] For the sake of transparency and replicability we took the list of Italian architects that features in Wikipedia, in its version from March 2023.

[16] As if Mattai de Rossi would have become CLIP's proxy for 'generic Romeness'.

Figure 6 Results of the multiscalar evaluation of Rome compared to 109 Italian provinces (left), 44 European capitals (centre) and 241 world capitals (right). Rome and the Vatican City are the only results reported in black.

4 Discussion

Our spatialized application of a CLIP model to urban images provides several intriguing initial results that demonstrate the promise of this approach while raising critical questions for further inquiry. Thematic mapping of model outputs reveals distinct spatial patterns of perceived urban qualities, suggesting CLIP embeddings encode certain logic around geographical variations in cityscape character. However, we observe potential gaps between computational assessments of Rome by OpenCLIP and established qualitative knowledge. If this stems from training data biases toward more contemporary materials and aesthetics, it remains to be tested. This underscores the need to critically examine how such systems conceptualise notions like cultural heritage and character.

The computational eye we put in place discerns patterns at scale that elude the human gaze, unveiling the essence of ordinary urbanism encoded across countless fragments. Results seem to suggest that the city becomes legible, and most importantly, comprehensible in its complexity not through spectacular skylines or iconic monuments, but through the constant landscape of mundane streets and urban tissue. This attempt at penetrating this generic realm underscores AI's capacity to potentially decode cities in the aggregate rather than as constellations of landmarks.

This computational discernment of the urban generic represents a major divergence from past lenses that privileged the monumental over the mundane. It also contrasts with critics like Lewis Mumford who decried "formless masses of urban residues" of modern cities lacking intentional design (Mumford, 1961). The capacities of general learner foundation models like CLIP and OpenCLIP, however, provide renewed attention to the cumulative legibility and qualities of ordinary urban fabrics. Their data-driven apprehension of the generic city promises new perspectives on urban form and experience. Through intentional, and critical application, multimodal models like CLIP may reveal new insights into ordinary urbanism, provided their limits are acknowledged.

In any case, a critical lens and contextualised interpretation are vital to unpack the cultural logics and biases encoded in AI’s worldviews. We therefore want to affirm approaching computational urbanism critically rather than reductively. Treating urban data as *transparent* risks a naive technocratic solutionist approach to complex phenomena that exceed computational capture. Critical perspectives are vital for examining what insights computational techniques provide, but also what they inevitably exclude. Multimodal Urban Studies offers an innovative and productive lens, but only one among many required. Urban research must retain pluralistic, contextualised approaches.

These types of models and their machinic view on our world, our cultures, and cities are pervasive and increasingly endowed with agency to shape our world. While we cannot emphasise enough the continued need for critical human interpretation, theorization and contextualization, we also advocate taking these technologies seriously, which not only add unprecedented synthetic analytical capacities but potentially subvert and expand the epistemological regimes of urban phenomena.

References:

Augé, M. (1995). Non-places: Introduction to an Anthropology of Supermodernity. Verso.
Batty, M. (2013). The new science of cities. MIT press.
Batty, M. (2013). Big data, smart cities and city planning. Dialogues in Human Geography, 3(3), 274-279
Bengio, Y., Courville, A., & Vincent, P. (2013). Representation Learning: A Review and New Perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798-1828.
Brenner, N. (2013). Theses on urbanization. Public culture, 25(1 69), 85-114.
Buchanan, I. (2000). Michel de Certeau, Cultural Theorist. Sage
DeVries, T., Misra, I., Wang, C., & van Merriënboer, B. (2022). PsyNeuLink: A framework for neural and cognitive modeling using Python. Computational Brain & Behavior, 5(1), 1-10
Eliade, M. (1957). The Sacred and the Profane: The Nature of Religion. Harcourt Brace.
Farocki, H. (2004). Phantom images. Public, 29(29), 12-24.
Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.
Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep Learning (Vol. 1). MIT Press Cambridge.
Greenfield, A. (2013). Against the smart city (The city is here for you to use Book 1). Do Projects.
Jacobs, J. (1961). The death and life of great American cities. Vintage
Kitchin, R. (2014). The real-time city? Big data and smart urbanism. GeoJournal, 79(1), 1-14
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.
Lefebvre, H. (2003). The urban revolution. U of Minnesota Press.
Manovich, L. (2017). AI aesthetics. Moscow: Strelka Press.
Massey, D. B. (2005). For space. Sage.
McCormack, J., Gifford, T., & Manning, P. (2019). Autonomy, Authenticity, Authorship and Intention in Computer Generated Art. ICCC, 196-203.
Naik, N., Philipoom, J., Raskar, R., & Hidalgo, C. (2014). Streetscore-predicting the perceived safety of one million streetscapes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 793-799).
Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (pp. 8748-8763.
Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).
Saxena, S., Tripathi, G., & Talukdar, P. P. (2022). Vokenization: Improving generalization of vision-and-language models via textual decomposition. arXiv preprint arXiv:2205.03750.
Tonkiss, F. (2013). Cities by Design: The Social Life of Urban Form. Cambridge: Polity Press.
Whyte, W. H. (1980). The social life of small urban spaces. Conservation Foundation.
Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. IEEE Computational Intelligence Magazine, 13(3), 55-75.
Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In European conference on computer vision (pp. 818-833). Springer, Cham.
Zheng, Y. T., Capra, L., Wolfson, O., & Yang, H. (2019). Urban Computing: Concepts, Methodologies, and Applications. ACM Transactions on Intelligent Systems and Technology (TIST), 5(3), 1-55.
Zhu, Y., Jiang, Z., Zhao, F., Terzopoulos, D., & Zhu, S. C. (2017). Inferring forces and learning human utilities from videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6234-6243).