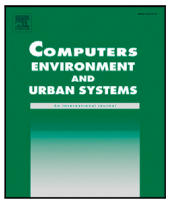
Computers, Environment and Urban Systems 125 (2026) 102393

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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



Learning street view representations based on a spatiotemporal contrastive learning framework

Yong Li [a,b,1], Yingjing Huang [a,1], Fan Zhang [a,*]

[a] Institute of Remote Sensing and Geographic Information System, School of Earth and Space Sciences, Peking University, Beijing, 100871, China
[b] Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong Special Administrative Region of China

ARTICLE INFO

Keywords:
Representation learning
Street view images
Urban analytics
Urban environment

ABSTRACT

Street view imagery has become an important data source for urban studies, supporting various urban tasks such as environmental perception and socioeconomic predictions. Classic methods predominantly rely on handcrafted features or supervised machine learning to derive information from the images. However, these methods often fail to capture the hierarchical semantics of urban environments: at the visual layer they cannot selectively represent dynamic versus static objects, while at the higher contextual layer they cannot abstract the collective ambience of a scene beyond tangible visual content, which in turn limits their effectiveness in tasks such as place recognition and socioeconomic inference. Essentially, this limitation arises because different urban tasks rely on fundamentally different invariances across space and time. To address this challenge, we propose the spatiotemporal contrastive learning framework, a novel self-supervised framework that systematically organizes representation learning for urban scenes. This framework defines distinct pre-training strategies by selectively contrasting what remains invariant versus what changes across the dimensions of space and time, enabling the model to isolate specific urban features like dynamic elements, static structures, or neighborhood ambiance. The validation experiments confirm that each contrastive strategy produces specialized representations that significantly outperform established baselines on their corresponding tasks. This study provides not only a novel representation framework but also a rigorous benchmark that enhances the applicability of visual data in urban science. The code is available at https://github.com/yonglleee/UrbanSTCL.

1. Introduction

With the growing availability of street view imagery (Naik et al., 2017; Zhang, Salazar-Miranda, et al., 2024), cities are leveraging large-scale visual data for diverse tasks such as place recognition (Lowry et al., 2015), urban perception analysis (Dubey et al., 2016; Zhang et al., 2018), road condition assessment (Chacra & Zelek, 2018), and socioeconomic prediction (Gebru et al., 2017; Wang, Li, & Rajagopal, 2020). Unlike classic object-centric vision tasks, these urban applications focus on distinct aspects of the urban environment. For instance, place recognition relies on invariant features including buildings and roads, while measuring human perceptions of a place relies on elements such as building conditions, street lighting, human activities, and vegetations to assess the overall perceptions within a scene, and socioeconomic prediction focuses on a spatial-invariant neighborhood atmosphere, capturing physical, social, cultural, and functional features across nearby areas. Learning effective street view representations that adapt to these varied needs, particularly in capturing both spatial and temporal dynamics of urban environments, remains a key challenge.

To address the challenge of learning such adaptable representations, researchers have increasingly turned to self-supervised learning (SSL). Self-supervised learning, leveraging techniques like contrastive learning (Chen et al., 2020, 2021; He et al., 2020) and masked modeling (He et al., 2022; Xie et al., 2022), has demonstrated outstanding performance in classical vision tasks such as image classification (Radford et al., 2021), object detection (He et al., 2022), and semantic segmentation (Wang, Zhang, et al., 2020), often surpassing traditional supervised learning approaches. However, current self-supervised methods tend to encode as much semantic and structural information as possible (Huang et al., 2024; Park et al., 2023), which does not fully align with the diverse requirements of urban tasks. For example, they may struggle to differentiate between the static features needed for place recognition (Lowry et al., 2015) and the dynamic elements critical for human perception of places (Dubey et al., 2016; Zhang et al., 2018), or to capture the spatial consistency required for socioeconomic prediction (Wang, Li, & Rajagopal, 2020).

* Corresponding author.
E-mail address: fanzhanggis@pku.edu.cn (F. Zhang).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.compenvurbsys.2025.102393
Received 22 September 2025; Received in revised form 10 December 2025; Accepted 12 December 2025

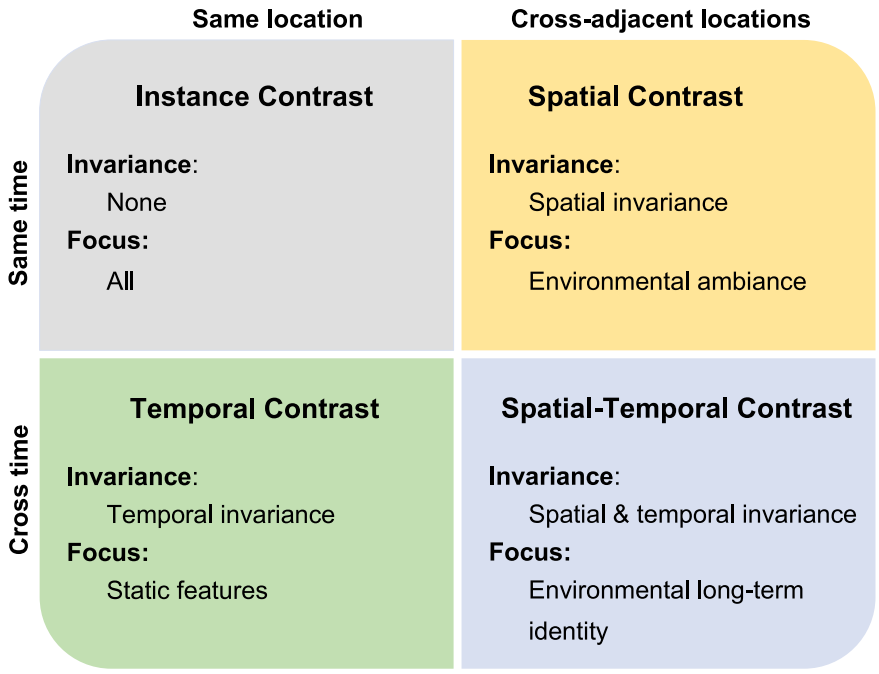


Fig. 1. The spatiotemporal contrastive learning framework. Our framework organizes four contrastive learning strategies based on their spatiotemporal context. The axes define the relationship between the two images in a positive pair. The vertical axis distinguishes between images captured at the same time versus those captured across time, while the horizontal axis distinguishes between images from the same location versus those across adjacent but different locations. This design allows each strategy to target a different type of invariance, yielding specialized representations with a distinct focus, as detailed in each quadrant.

In image representation learning, selectively encoding dynamic and static information in urban environments and the ambiance they create is highly important but inherently challenging (Cordts et al., 2016). Achieving precise encoding of such information typically requires separately labeling dynamic and static elements and using specific training strategies (Cheng et al., 2017; Wang et al., 2019) (e.g., masking dynamic elements when encoding static ones). However, both the labeling and training processes are fraught with difficulties. Factors such as lighting conditions, vegetation appearance, and ground litter are challenging to label objectively and consistently. This makes it nearly impossible to accurately represent these complex environmental factors using traditional datasets (e.g., ImageNet (Deng et al., 2009), Places (Zhou et al., 2017)) and classical methods (supervised or self-supervised).

To address these challenges, we propose a contrastive street-view representation learning framework that explicitly leverages timestamp and geolocations—types of metadata largely absent from standard image datasets. The core idea is to form complementary positive pairs that target different invariances: (i) Temporal Contrast — positive pairs formed from the same location captured at different times — drive the encoder to emphasize time-invariant, static attributes of the built environment (e.g., buildings, infrastructure) and to suppress sensitivity to dynamic elements (pedestrians, vehicles), benefiting tasks such as place recognition. (ii) Spatial Contrast — positive pairs formed from images taken at the same time and adjacent but different locations — encourage representations that are stable within an urban neighborhood, capturing its socioeconomic "ambience" while reducing sensitivity to object-level variations, which supports neighborhood-scale socioeconomic estimation. (iii) Instance Contrast essentially reduces to classical instance-level contrastive learning, yielding representations that preserve the full scene (both static and dynamic content, as well as overall ambience) for human-perception-oriented tasks. (iv) Spatial–temporal Contrast — positive pairs spanning both temporal and spatial variations promote invariance over both space and time, capturing more enduring, higher-level characteristics — such as historical and cultural character—that support related urban analytics.

We validate the effectiveness of our primary hypotheses (Instance, Spatial, and Temporal contrast) across multiple urban tasks. While our framework also conceptualizes a Spatial–temporal contrast for learning deep historical and cultural patterns, we leave its experimental validation for future work, given the difficulty of collecting the necessary ground-truth data for its corresponding downstream tasks. Experimental results demonstrate that different contrastive learning objectives can learn different types of features that are more suitable for their respective urban tasks. We also conduct an in-depth analysis of the reasons behind the performance of different contrastive methods, further underscoring the importance of targeted learning strategies. This study systematically explores representation learning strategies in urban studies based on street view images, provides a valuable benchmark, and enhances the applicability of visual data in urban science.

2. Related work

2.1. Street view representation learning for urban tasks

Street view imagery has been widely used in various urban tasks (Gebru et al., 2017; Naik et al., 2017), such as road defect detection (Chacra & Zelek, 2018), traffic prediction (Zhang, Li, & Zhang, 2024), urban function recognition (Huang et al., 2023), and socioeconomic prediction (Fan et al., 2023). However, existing research on street view representation often relies on supervised models trained on datasets like Places365 (Zhou et al., 2017) or directly uses the pixel proportions of semantic segmentation results. These approaches fail to fully capture the rich semantic information embedded in street view imagery. Unlike natural images, street view imagery not only contains complex visual semantics but also encodes valuable spatiotemporal information in its metadata. Effectively representing this dual semantic nature — both visual and spatiotemporal — remains a significant challenge for improving its use in urban tasks. Although a few studies have explored spatiotemporal self-supervised learning approaches to represent street view imagery (Stalder et al., 2024), these methods

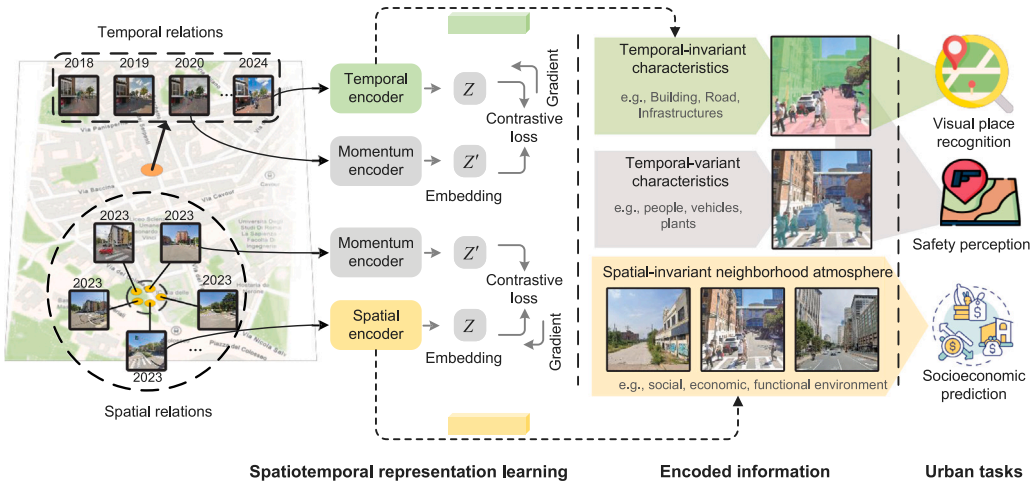


Fig. 2. Spatiotemporal contrastive learning with street view images for diverse urban tasks. Temporal relations are constructed by capturing images from the same location at different times (e.g., 2018–2024), while spatial relations are established using nearby images taken at the same time. The temporal contrast captures temporal-invariant features (e.g., buildings, roads, infrastructure), while the spatial contrast captures spatial-invariant neighborhood atmosphere, reflecting the physical, social, and cultural environment. Different representation learning strategies are designed to support various urban tasks, such as visual place recognition, safety perception, and socioeconomic prediction.

fail to explore the natural meanings of the spatiotemporal attributes of street view imagery and how to leverage these attributes to construct self-supervised methods suitable for various urban tasks. For instance, Urban2Vec (Wang, Li, & Rajagopal, 2020) incorporates spatial information into self-supervised training by constructing positive sample pairs based on nearest neighbors, while KnowCL (Liu et al., 2023) integrates knowledge graphs with contrastive learning to align locale and visual semantics, improving the accuracy of socioeconomic prediction using street view imagery.

2.2. Self-supervised representation learning for images

Self-supervised learning (SSL) aims to leverage large amounts of unlabeled data to learn effective feature representations by designing proxy tasks based on the inherent structure of the data itself. This approach reduces reliance on manually annotated datasets, making it a powerful paradigm for representation learning. In computer vision, contrastive learning stands out as one of the most widely adopted SSL methods. These methods train models using discriminative pretext tasks, with the core idea of learning robust data representations by distinguishing between different samples (Caron et al., 2021; Chen & He, 2021; Wu et al., 2018). Notable examples of contrastive learning algorithms include SimCLR, MoCo, and BYOL (Chen et al., 2020, 2021; Grill et al., 2020; He et al., 2020). While these approaches have achieved significant success, they predominantly focus on natural images lacking spatiotemporal context, often targeting classic computer vision tasks such as semantic segmentation and object detection. These tasks require encoding as much information as possible from static images. However, urban tasks involving street view imagery present distinct challenges, where spatial and temporal dependencies play a critical role in capturing the dynamics of urban environments. To address these challenges, spatiotemporal self-supervised learning extends traditional SSL methods by incorporating temporal coherence (Manas et al., 2021; van den Oord et al., 2019) and geographic context (Ayush et al., 2021; Deuser et al., 2023; Guo et al., 2024; Klemmer et al., 2024; Mai et al., 2023). These adaptations have proven effective in domains like remote sensing and multi-view learning, yet their application to street view imagery remains underexplored. A more integrated spatiotemporal self-supervised framework is essential to better model the dynamic nature of urban landscapes and enhance the performance of urban-related applications.

3. Learning street view representations with spatiotemporal contrast

Our approach to learning urban representations is guided by the spatiotemporal contrastive learning framework (Fig. 1), a unified framework designed to leverage the unique attributes of street view imagery. This framework organizes representation learning along two fundamental axes that define how positive pairs are constructed: the spatial axis, which considers whether pairs are from the same place or different locations within a neighborhood, and the time axis, which considers whether they are from the same time. This creates four distinct quadrants of contrastive learning, each designed to isolate a specific type of urban feature: Instance Contrast, Temporal Contrast, Spatial Contrast, and the conceptual Spatial–temporal Contrast. It is important to note that GSV-self, GSV-spatial, and GSV-temporal represent distinct contrastive learning objectives used to train the model. This section details the specific implementation of these learning strategies (Fig. 2).

3.1. Instance contrast learning

Instance Contrast Learning serves as the foundational strategy in our framework, designed to extract robust features from individual street view images. This approach is built on the core principle of contrastive learning: learning representations by minimizing the distance between positive samples and maximizing the distance from negative samples in a feature space. Crucially, in Instance Contrast, a positive pair is generated by applying two different random augmentations (e.g., cropping, color jitter) to the same source image, creating two distinct but semantically identical views. All other images in a batch are treated as negative samples.

By optimizing the InfoNCE loss function, the model learns to reduce the distance between positive pairs in the feature space and increase the distance from negative samples, thus improving the feature representation learning.

To learn augmentation invariance, we define the instance contrastive loss. Given a positive pair of augmented views (x_i, x_j) derived from the same source image, the instance contrastive loss is:

$$\mathcal{L}\mathrm{i} = -\log \frac{\exp\left(x_i \cdot x_j/\tau\right)}{\exp\left(x_i \cdot x_j/\tau\right) + \sum x_k^- \exp\left(x_i \cdot x_k^-/\tau\right)} \tag{1}$$

where x_i and x_j are the feature representations of the positive augmented views, and x_k^- represents the negative samples. This loss encourages the model to maximize the similarity between different views of the same image while minimizing their similarity to all other images. Building on this contrastive learning framework, we introduce temporal and spatial contrasts for constructing positive pairs from street view images.

To enable the use of a large and consistent dictionary of negative samples without the need for massive batch sizes, we adopt a momentum encoder framework (He et al., 2020). In this approach, the key representations (k^+ and k^-) are generated by a separate momentum encoder. Crucially, this encoder is not updated through backpropagation, which prevents the dictionary keys from becoming inconsistent as the model trains. Instead, its parameters (θ_k) are a slowly evolving exponential moving average of the query encoder's parameters (θ_q):

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q$$

With a high momentum coefficient m (e.g., 0.999), this method ensures that the keys in our dynamic dictionary remain consistent, providing a stable target that is essential for effective contrastive learning. The following sections introduce our novel temporal and spatial contrastive strategies. They are built upon this same foundational architecture — using InfoNCE loss and a momentum encoder — but critically redefine the method for constructing positive pairs to capture specific spatiotemporal features.

3.2. Temporal contrastive learning

Street view images captured at the same location but at different times differ from video frames because the intervals between shots are not fixed. Unlike remote sensing images, street view images taken at different times are not perfectly aligned in terms of geographic locations. Due to the typical spatial and angular shifts between images captured at different times, we define positive temporal pairs based on their close proximity — such as being taken just a short distance apart — and having the same shooting angle, ensuring sufficient consistency without demanding exact alignment. The historical street view image set for each location can be represented as $T = [t_1, t_2, \ldots, t_n]$, where t_i denotes the images captured at different times. Since the number of images varies for each location, resulting in different values of n, we randomly selected two images from different time periods within each set to serve as a positive pair. The aim of temporal contrast is to capture the invariant features of the same location over time. This means that even though the images are taken at different times, the model should learn to recognize the consistent characteristics of the scene.

To capture invariant features of the same location over time, we define the temporal contrastive loss. Given a positive sample pair (t_i, t_j) that meets temporal conditions ((images taken in close proximity and from the same angle)), the temporal contrastive loss is:

$$\mathcal{L}_{\mathrm{t}} = -\log \frac{\exp\left(t_i \cdot t_j/\tau\right)}{\exp\left(t_i \cdot t_j/\tau\right) + \sum_{t_k^-} \exp\left(t_i \cdot t_k^-/\tau\right)} \tag{2}$$

where t_i and t_j are feature representations of the positive temporal samples, t_k^- denotes negative samples from different locations or angles, and τ is the temperature parameter for scaling. This formulation aims to maximize similarity between the same location's images taken at different times while minimizing similarity to negatives.

3.3. Spatial contrastive learning

Capturing the spatial consistency of an urban area is essential for accurately representing the urban physical environment. Spatial consistency refers to the ability to recognize that different locations within the same urban area still represent the same underlying physical characteristics. To achieve this, we treat all street view images captured within a specific urban area as sharing a common set of environmental characteristics, even if these images are taken from different angles or slightly different positions. This approach allows the model to account for variations in location while preserving the overall ambiance of the area. The set of street view images for a given urban area can be denoted as $S = \{s_1, s_2, \ldots, s_n\}$, where each s_i represents an image captured within the defined area. These images collectively provide a comprehensive spatial representation of the urban environment. We randomly select two samples (s_i, s_j) from the set S and treat them as positive pairs. This encourages the model to learn that despite slight variations in shooting angle or position, the images are part of the same spatial context.

To capture spatial consistency within an urban area, we define the spatial contrastive loss. Given a set of street view images $S = \{s_1, s_2, \ldots, s_n\}$ from the same urban area, we randomly select two samples (s_i, s_j) as a positive pair and define the spatial contrastive loss as:

$$\mathcal{L}_{\mathrm{s}} = -\log \frac{\exp\left(s_i \cdot s_j/\tau\right)}{\exp\left(s_i \cdot s_j/\tau\right) + \sum_{s_k^-} \exp\left(s_i \cdot s_k^-/\tau\right)} \tag{3}$$

where s_i and s_j are feature representations of the positive spatial samples, and s_k^- represents negative samples from different urban areas. This loss encourages the model to maximize similarity between images in the same urban area while minimizing similarity to negatives from other areas. By doing so, we enable the model to learn consistent and representative spatial features across the entire urban area.

3.4. Spatial–temporal contrastive learning

The fourth quadrant of our framework, Spatial–temporal Contrast, involves constructing positive pairs from images of different nearby locations taken at different times. The objective is to learn spatial–temporal invariance, forcing the model to discover the long-term core identity of a neighborhood. By filtering out both short-term temporal dynamics and hyperlocal spatial details, this representation would theoretically capture the enduring architectural character and functional essence of a region.

Formally, given a positive spatial–temporal pair (st_i, st_j), the loss function would be defined as:

$$\mathcal{L}_{\mathrm{st}} = -\log \frac{\exp\left(st_i \cdot st_j/\tau\right)}{\exp\left(st_i \cdot st_j/\tau\right) + \sum_{st_k^-} \exp\left(st_i \cdot st_k^-/\tau\right)} \tag{4}$$

where st_i and st_j are the feature representations of the positive pair, and st_k^- denotes negative samples from unrelated regions or time periods.

While conceptually powerful, we did not experimentally implement spatial–temporal contrast in this work. The primary challenge lies in identifying suitable downstream tasks and corresponding benchmark datasets for validation. Tasks that would benefit from such a representation, like analyzing the long-term evolution of urban fabric, require large-scale, longitudinal data that is often not readily available for standardized evaluation. Therefore, we posit this strategy as a promising and significant direction for future research, which will build upon the foundational work presented here.

4. Applying task-centric representations to urban applications

Urban environments exhibit both spatial and temporal complexities —locations change over time yet retain inherent characteristics, and different areas share structural similarities while maintaining distinct identities. Capturing these dynamics is essential for understanding cities, making tasks such as visual place recognition, socioeconomic prediction, and safety perception natural benchmarks for evaluating our contrastive learning framework. We assess their effectiveness across these urban tasks by pre-training models using self-supervised learning on temporal and spatial contrastive datasets. We also analyze how different contrastive strategies influence learned urban representations.

4.1. Urban tasks description

Understanding urban environments involves recognizing locations under varying conditions, inferring socioeconomic patterns from visual cues, and assessing perceived safety. Each of these tasks inherently involves distinct spatial and temporal challenges that align with our contrastive learning objectives. These three urban tasks — visual place recognition, socioeconomic prediction, and safety perception — collectively test a model's ability to disentangle invariant and dynamic urban features, as detailed below.

Visual place recognition. Locations undergo seasonal changes, construction, and variations in lighting, yet key structural elements remain invariant. The challenge in visual place recognition is to distinguish locations while being robust to these transient variations. A model that captures invariant features while ignoring irrelevant fluctuations improves visual place recognition performance. We evaluate visual place recognition performance using multiple datasets that capture diverse environmental variations. The CrossSeason dataset (Mans Larsson et al., 2019) focuses on seasonal changes, testing model robustness to variations in snow, foliage, and lighting throughout the year. The ESSEX dataset (Zaffar et al., 2021) introduces viewpoint and lighting diversity in urban and suburban settings, challenging the model's ability to recognize places under different perspectives. The Pittsburgh dataset (Arandjelović et al., 2018) extends this by incorporating large-scale street view imagery from Pittsburgh, supporting localization and geographic recognition. The SPED dataset (Chen et al., 2018) emphasizes temporal changes, containing images captured at different times to study scene dynamics and urban transformation. Lastly, the MapillarySLS dataset (Warburg et al., 2020) provides a globally distributed dataset with diverse street view images, aiding in tasks such as autonomous driving and broad-scale visual recognition. Together, these datasets comprehensively evaluate the model's ability to handle spatial, temporal, and environmental variations in visual place recognition.

Socioeconomic prediction. The urban environment reflects socioeconomic characteristics through its visual features, from infrastructure quality to commercial density. Inferring socioeconomic indicators requires recognizing patterns that extend beyond individual images to the broader urban context. A model that associates images from the same area while distinguishing them from those in different socioeconomic conditions provides stronger predictive capability. In our urban task, we used socioeconomic indicators provided by Fan et al. (2023), which include data from seven major metropolitan areas in the United States. The socioeconomic indicators cover various topics relevant to urban studies and are detailed in Table 1.

Safety perception. Perceived safety is influenced by multiple visual factors, including lighting, greenery, building conditions, and the openness of spaces, which vary across both space and time. A robust safety perception model could capture safety-related features while adapting to temporal changes caused by urban development or daily cycles. To evaluate our approach, we use PlacePulse 2.0 (Dubey et al., 2016), a large-scale dataset containing crowdsourced safety perception ratings for urban scenes. This dataset provides a diverse range of environments, enabling models to learn and generalize safety-related visual cues across different geographic and temporal contexts.

These urban tasks naturally reflect the challenges of disentangling invariant characteristics from dynamic variations, a fundamental objective in learning urban representations. Evaluating our contrastive learning models on these benchmarks allows us to assess their ability to capture meaningful urban features that generalize across different environments.

4.2. Street view data and pre-training datasets

We collect street view imagery to develop pre-training datasets for self-supervised learning models targeting urban tasks. Then, we apply our spatiotemporal contrastive framework to pre-train models, effectively capturing urban characteristics.

4.2.1. Data collection and preprocessing

To obtain street view imagery for both self-supervised model training and socioeconomic prediction, we first sourced road network data for each city using the OSMnx library (Boeing, 2017) from OpenStreetMap. We then generated query points along these road networks at regular intervals of 15 m. The Google Street View (GSV) Application Programming Interface (API) was subsequently utilized to retrieve and download street view images.

Since the visual place recognition and safety perception datasets include a wide range of street view images from different cities, while the socioeconomic prediction task focuses more on local city characteristics, we constructed two separate datasets — a global version and a local version — for testing on different urban tasks. For the global version, to capture a broad spectrum of urban environments, we trained our self-supervised models on data collected from ten diverse and representative global cities including Amsterdam, Barcelona, Boston Metropolitan Area (Boston), Buenos Aires, Dubai–Sharjah (Dubai), Johannesburg, Los Angeles, Melbourne, Seoul, and Singapore. These cities were carefully selected to encompass a variety of geographical locations, cultural backgrounds, and urban forms, ensuring the diversity and richness of our training dataset. We collected historical images of ten global cities from the GSV API, which resulted in a total of over 42 million street view images used for pre-training. For the local version, we selected street view images from Los Angeles to construct different contrastive datasets tailored to the specific needs of the socioeconomic prediction task in that city. The construction methods of datasets are similar to the global version.

4.2.2. Pre-training datasets construction

Based on the street view pre-training datasets, we constructed three distinct contrastive datasets corresponding to different contrastive learning models for both global and local versions: instance contrast, temporal contrastive, and spatial contrastive datasets. To benchmark against the MoCov3 baseline trained on ImageNet, each dataset was standardized to consist of 1 million image pairs. This uniform dataset size facilitates a fair comparison among the models by ensuring that each receives an equal amount of training data.

Instance contrast dataset. For the instance contrast dataset, we randomly selected 100,000 images from each of the 10 cities, resulting in a total of 1 million images. Positive pairs were generated during training by applying data augmentation techniques to these images, following the settings used in MoCo v3 (Chen et al., 2021). Additionally, for the local version, we constructed an instance contrast dataset based solely on Los Angeles using the same method.

Temporal contrastive dataset. In constructing the temporal contrastive dataset, we randomly selected 100,000 street view sampling points from each of the 10 cities, totaling 1 million sampling points. At each sampling point, we retrieved images taken at different times but in close proximity, specifically within 5 m, and from the same shooting angle. This constraint ensures that the images remain spatially and visually consistent despite temporal differences, minimizing the impact of significant positional or perspective shifts on the temporal contrastive learning process. Two images were randomly selected from the temporal sequence to form a positive pair, resulting in 1 million temporal positive pairs. Similarly to the instance contrast dataset, we constructed an additional temporal contrastive dataset based solely on Los Angeles using the same method.

Spatial contrastive dataset. For the global spatial contrastive dataset, we defined a 100-meter buffer zone as the spatial unit for contrastive analysis. This 100-meter radius was selected to provide a standardized spatial scale across different countries, as national census units vary significantly in size and definition, necessitating a uniform buffer for consistent global comparisons. From each buffer zone, we randomly selected two images to form positive pairs, and out of all the spatial positive pairs generated, we then randomly selected 1 million pairs to create the spatial contrastive dataset. Notably, we did not

Table 1
Socioeconomic Indicators for Urban Prediction: Crime, Health, Poverty, and Transport Metrics by Spatial Unit.

Topic	Indicator	Label
Crime	Violent crime occurrence per spatial unit	Log(Violent Crime)
	Violent theft-related crime occurrence per spatial unit	Log(Petty Crime)
Health	Model-based estimate for crude prevalence of cancer (excluding skin cancer) among adults aged ≥ 18 years	% Cancer Health
	Model-based estimate for crude prevalence of diagnosed diabetes among adults aged ≥ 18 years	% Diabetes
	Model-based estimate for crude prevalence of no leisure-time physical activity among adults aged ≥ 18 years	% LPA
	Model-based estimate for crude prevalence of mental health not good for ≥ 14 days among adults aged ≥ 18 years	% Mental Health
	Model-based estimate for crude prevalence of obesity among adults aged ≥ 18 years	% Obesity
	Model-based estimate for crude prevalence of physical health not good for ≥ 14 days among adults aged ≥ 18 years	% Physical Health
Poverty	Median Household Income	Log(Income)
	% Individuals with poverty status determined: below 100% poverty line	% Poverty Line (100%)
	% Individuals with poverty status determined: below 200% poverty line	% Poverty Line (200%)
Transport	% Population (>16) commute by driving alone	% Drive Alone
	Estimated personal miles traveled on a working weekday	PMT
	Estimated personal trips traveled on a working weekday	PTRP
	Estimated vehicle miles traveled on a working weekday	VMT
	Estimated vehicle trips traveled on a working weekday	VTRP
	% Population (>16) commute by public transit	%Public Transit
	% Population (>16) commute by walking and biking	%Walk

impose restrictions on the shooting angle for these positive pairs, allowing the model to focus on the broader urban environment rather than specific street layouts. In contrast, for the local version, we adopted U.S. census block groups (CBGs) as the spatial units for contrastive analysis, leveraging their standardized definition across the United States. Positive pairs were constructed based on the boundaries of these CBGs, aligning the spatial scale with the U.S.-specific socioeconomic dataset.

4.2.3. Pre-training details

We use AdamW (Loshchilov & Hutter, 2019) as the optimizer, a common choice for training ViT base (Dosovitskiy et al., 2021) models, with a weight decay of 1e−6. For each dataset, we use a mini-batch size of 1024 and an initial learning rate of 6e−6. The model is trained for 300 epochs, starting with a 40 epoch warmup (Goyal et al., 2018), followed by a cosine decay schedule for learning rate decay (Loshchilov & Hutter, 2017). Training the ViT Base model for 300 epochs on 4 Nvidia A800 GPUs takes approximately 71 h.

5. Results

We evaluate our models on three tasks — visual place recognition, socioeconomic prediction, and safety perception — each aligning with different contrastive learning strategies. Visual place recognition benefits from temporal contrastive learning to enhance stability across time. Socioeconomic prediction relies on spatial contrastive learning to capture neighborhood patterns. Safety perception leverages instance contrast learning to extract global scene features. These contrastive objectives enable our model to learn robust and transferable urban representations.

To demonstrate the effectiveness of our proposed methods, we benchmark their performance against several comparators. Our baselines include Urban2Vec (Wang, Li, & Rajagopal, 2020), a prominent urban representation method that we retrained on our Spatial contrastive dataset for a fair comparison, where we exclusively employed the visual component and excluded the POI module to ensure a direct evaluation of street view representations; Urban2Vec-ViT, a variant where we replaced the original Inception-V3 backbone with ViT-Base to ensure architectural consistency with our proposed models; and ImageNet-self, a model pre-trained on ImageNet with the MoCo-v3 self-supervised method to test the transferability of general visual features (Chen et al., 2021). In addition, we establish a direct baseline, GSV-self, using instance-level contrastive learning on our Google Street View dataset. We evaluate these models against our proposed GSV-spatial and GSV-temporal approaches, which explicitly learn from spatial and temporal relationships, respectively. This comprehensive comparison is designed to highlight the unique advantages of incorporating spatiotemporal context into representation learning for urban environments.

5.1. Visual place recognition

Visual place recognition is a crucial urban task that aims to identify specific locations based on visual input. This task requires the removal of temporal disturbances to focus on invariant information that does not change over time, demanding feature extraction that effectively distinguishes constant characteristics in the environment to improve recognition accuracy.

To evaluate the model's performance in visual place recognition tasks, we used several benchmark datasets: CrossSeason (Mans Larsson et al., 2019), Essex (Zaffar et al., 2021), Pitts250k, Pitts30k (Arandjelović et al., 2018), SPED (Chen et al., 2018), and MapillarySLS (Warburg et al., 2020) datasets. Detailed information about these benchmark datasets is described in Section 4.1. We formulate VPR as a large-scale image retrieval task. For each dataset, the test images serve as "queries", which are used to search against a large, geotagged database of images. The objective is to retrieve the database images that are geographically closest to the true location of the query image.

The models were tested by freezing the backbone of the pre-trained ViT and extracting the [CLS] token for visual place recognition tasks. We assessed performance using the Recall@K metric, measuring the models' ability to correctly identify query image locations among the top-k most similar database images.

The critical test for temporal and environmental invariance lies in how these query and database sets are constructed: images of the same location are deliberately captured under different conditions. For instance, in the CrossSeason dataset, the database may consist of images collected in summer, while the corresponding queries are the same

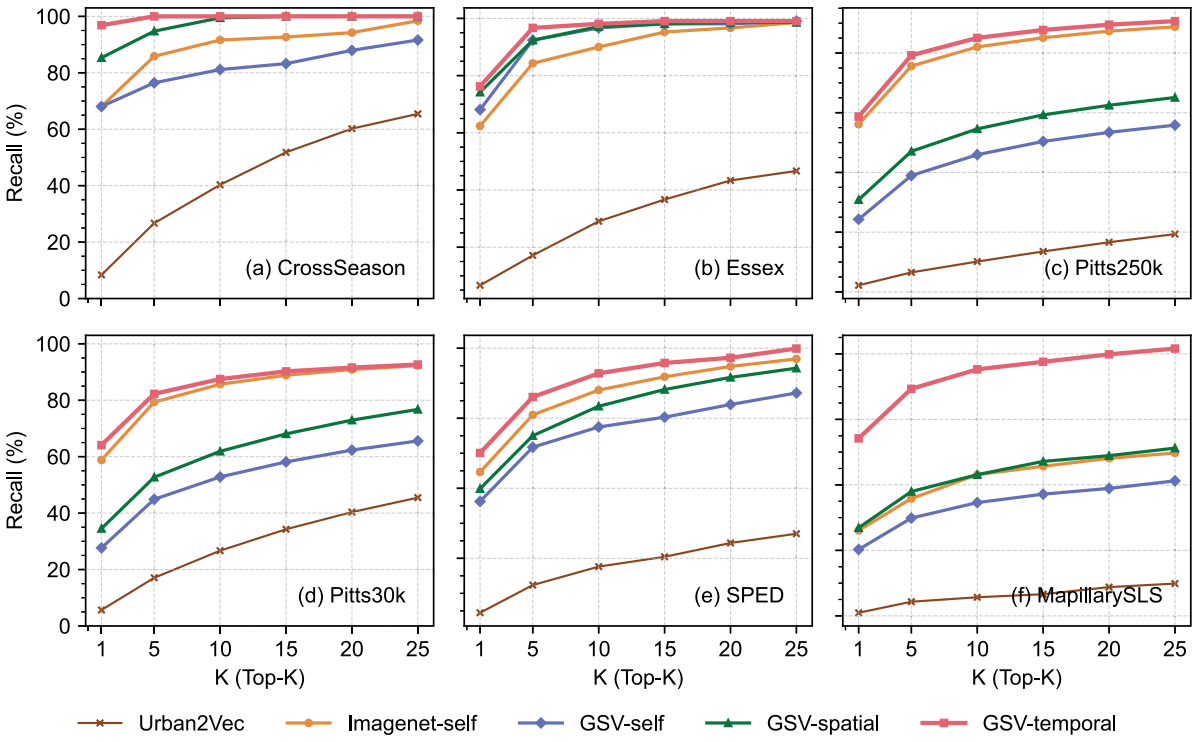


Fig. 3. Performance comparison on different visual place recognition datasets (Recall@K in %).

locations captured in winter. A high Recall@K score in this scenario indicates that the model has successfully learned a representation that is robust to drastic seasonal changes (e.g., snow, foliage, lighting) and focuses on the underlying, invariant structural features of the place. This same principle of mismatched conditions is applied in other datasets to test for different invariances, such as long-term changes in SPED and viewpoint shifts in ESSEX.

In Fig. 3, the GSV-temporal model demonstrates exceptional performance on the CrossSeason dataset, achieving a recall value of 100% across all K values. This indicates its robust capability in cross-season visual place recognition tasks. In contrast, GSV-self and ImageNet-self exhibit significantly lower performance, suggesting their inability to effectively capture temporal invariant features. On the Essex dataset, GSV-temporal maintains a recall value exceeding 75%, with values of 99.05% for both K = 20 and K = 25. This highlights its insensitivity to dynamic changes in the environment, which allows it to outperform other models in this context. In the Pitts250k dataset, GSV-temporal consistently outperforms GSV-self and ImageNet-self in recall values, the GSV-temporal model also excels on the Pitts30k dataset, achieving a recall value of 90.23% at K = 15. underscoring its suitability for complex urban environments in visual place recognition tasks. For the SPED dataset, GSV-temporal displays superior recall values compared to other models, particularly with a notable performance at K = 5. In the MapillarySLS dataset, GSV-temporal showcases its outstanding performance again, with a recall value of 77.57% at K = 15.

In summary, the GSV-temporal model consistently outperforms other models across multiple datasets, particularly in visual place recognition tasks. Its insensitivity to temporal and environmental changes positions it as a superior choice for this application, revealing significant potential for practical use.

5.2. Socioeconomic prediction

The socioeconomic prediction task uses street view images to infer the socioeconomic status of urban areas. It emphasizes learning the overall ambiance of a region rather than specific geometric features, highlighting the need for feature extraction to focus on similarities between regions to understand economic conditions and developmental dynamics better.

In the urban task of predicting socioeconomic indicators, we utilized the socioeconomic dataset published by Fan et al. (2023), which contains 18 socioeconomic indicators across seven major cities in the United States (Table 1). We take the socioeconomic prediction of Los Angeles as an example. Detailed descriptions are provided in Section 4.1. We first extracted street view features from the images using the pre-trained models of the local version. These features were then aggregated using the mean values at the block group level. The aggregated features were used as input features to predict socioeconomic indicators for each block group.

For prediction model training and evaluation, we split each city's dataset into a training set (70%) and a testing set (30%). We used LASSO (Tibshirani, 1996) as the regressor to evaluate the predictive performance of the image features extracted by the different pre-trained models. Additionally, we applied 5-fold cross-validation to ensure robust evaluation. This approach allows for a fair comparison of the different contrastive learning models in capturing visual features that are meaningfully correlated with socioeconomic indicators.

The results of socioeconomic predictions are shown in Table 2. Overall, models pre-trained on street view images significantly outperform the model pre-trained on the ImageNet dataset. Specifically, across all 18 indicators, the model pre-trained on the general ImageNet dataset achieved an average R^2 of 0.5209. while the prominent street-view-based method Urban2Vec scored 0.3464. In contrast, models on street view images achieved average R^2 scores of 0.5609 for instance contrast, 0.5714 for temporal contrastive, and 0.5888 for spatial contrastive models, respectively. Furthermore, both temporal and spatial contrastive pre-training models capture more socioeconomic-related information compared to the instance contrast approach, with spatial contrastive demonstrating the highest performance. This trend is consistent across most socioeconomic indicators, showing the strongest predictive performance for Health-related indicators and the least for Crime-related indicators.

Table 2
Model performance comparison on socioeconomic prediction tasks based on LASSO across contrastive models.

Topic	Label	Urban2Vec	Urban2Vec-ViT	GSV-self	GSV-spatial	GSV-temporal	ImageNet-self
Crime	Log(Violent Crime)	0.2369	0.3139	0.4203	**0.4287**	0.4194	0.4146
	Log(Petty Crime)	0.0634	0.1179	0.1810	0.1877	**0.1892**	0.1667
	Total	0.1501	0.2159	0.3007	**0.3082**	0.3043	0.2906
Health	% Cancer Health	0.3275	0.4216	0.6644	**0.6969**	0.6618	0.6053
	% Diabetes	0.3158	0.422	0.6589	**0.6942**	0.6796	0.6172
	% LPA	0.4458	0.5708	0.8001	**0.8337**	0.8221	0.7671
	% Mental Health	0.4206	0.4749	0.7088	**0.7510**	0.7291	0.6753
	% Obesity	0.3861	0.4432	0.7628	**0.7886**	0.7797	0.7175
	% Physical Health	0.3980	0.4497	0.7120	**0.7399**	0.7314	0.6752
	Total	0.3823	0.4637	0.7178	**0.7507**	0.7340	0.6763
Poverty	Log(Income)	0.3278	0.4438	0.6561	**0.6816**	0.6735	0.6096
	% Poverty Line (100%)	0.1246	0.1567	0.1948	**0.2227**	0.1833	0.1718
	% Poverty Line (200%)	0.3469	0.4454	0.6154	0.6377	**0.6401**	0.5893
	Total	0.2663	0.3486	0.4888	**0.5140**	0.4990	0.4569
Transport	% Drive Alone	0.1765	0.2591	0.3841	**0.3991**	0.3835	0.3582
	PMT	0.2974	0.4349	0.6196	**0.6447**	0.6289	0.5379
	PTRP	0.3269	0.4168	0.6024	**0.6385**	0.6087	0.5302
	VMT	0.4072	0.4645	0.6647	**0.6921**	0.6874	0.6163
	VTRP	0.3741	0.4888	0.6900	**0.6994**	0.6991	0.6436
	%Public Transit	0.1162	0.4237	0.5226	**0.5700**	0.5339	0.4726
	%Walk	0.2636	0.2551	0.2383	**0.2925**	0.2340	0.2080
	Total	0.2803	0.3918	0.5317	**0.5623**	0.5394	0.4810
Overall Total		0.3464	0.3890	0.5609	**0.5888**	0.5714	0.5209

Table 3
Evaluation metrics of different models on the safety perception classification task.

Model	Accuracy (%)	Recall (%)	F1 Score (%)	AUC Score (%)
Urban2Vec	79.25	64.52	69.44	76.12
Urban2Vec-ViT	80.94	67.58	70.54	78.11
ImageNet-self	83.25	70.32	75.43	80.51
GSV-temporal	84.91	65.16	75.94	80.72
GSV-spatial	86.08	68.39	78.23	82.33
GSV-self	**88.68**	**77.42**	**83.33**	**86.29**

These findings suggest that spatial contrastive pre-training effectively captures the overall ambiance of urban areas, enabling more precise predictions of regional socioeconomic information. Additionally, temporal contrastive pre-training filters out random factors and dynamic elements in the images, enhancing the reliability of socioeconomic predictions.

5.3. Safety perception

The safety perception task involves using street view imagery to estimate how safe people perceive a given scene to be. To make accurate estimates, this task requires analyzing all relevant elements within the scene, as each can contribute to the overall perception of safety, particularly elements such as trees and vehicles (Zhang et al., 2018).

We selected the PlacePlus 2.0 (Dubey et al., 2016) dataset for the urban task of human environmental perception, filtering out over 1144 images with safety perception scores below 3.5 and above 6.5, with 80% of the data used for training and 20% for testing. The model was trained using a linear binary classification approach for 20 epochs to effectively distinguish between low and high safety perception environments.

Table 3 compares the performance of various models in classifying safety perception in urban environments. As a baseline, the Urban2Vec model achieved an accuracy of 79.25% and an F1 score of 69.44%. In an improvement over this and all other methods, the GSV-self model achieved the highest performance, with a top accuracy of 88.68% and recall of 77.42%. This demonstrates its effectiveness in identifying both safe and unsafe environments while minimizing false negatives. Its F1 score of 83.33% indicates a balance between precision and recall, and the AUC score of 86.29% further confirms its ability to distinguish between safety levels across thresholds. Overall, the GSV-self model outperforms the others in all metrics, underscoring the strength of instance-level contrastive learning for urban safety perception tasks.

6. Discussion

We conduct interpretability analyses on the features learned by the different contrastive models to gain a deeper understanding of the information the models focus on and how this impacts performance on urban tasks.

6.1. Analysis of differences in spatiotemporal contrastive features

This section explores the differences in feature representation and retrieval tasks by comparing the Instance, Spatial, and Temporal contrastive methods. We use street view images from Chicago, analyzing the performance of these three contrastive methods in pre-trained models. For each model pre-trained with a unique contrastive learning objective, we extract a 768-dimensional feature vector representing the characteristics of street view images. In our experiment, to comprehensively evaluate the retrieval performance of these methods, we randomly selected 500 street view images from different locations in Chicago as query images, ensuring that each image originated from a distinct spatial location. For each query image, we used the Nearest Neighbors method in feature space to retrieve the top five street view images with the closest Cosine distance. This process generated a total of 500 sets of query and retrieval pairs, with five results for each query. Specifically, we used the Euclidean distance as a similarity measure to rank and obtain the top five retrieval results.

Finally, we randomly selected one set of queries and retrieval results for visualization. By comparing the year, heading, and feature distance of the retrieved results with the query image, we visually demonstrated the significant differences in retrieval characteristics among the three contrastive methods. Fig. 4 shows the retrieval results for a given query street view image using GSV-self, GSV-spatial, and GSV-temporal contrastive methods. On the left, the query image is displayed, including information about the year of capture (June 2018), heading (90°), geographic location (42.3951, 71.1217), and city (Chicago). The retrieval results are arranged in three rows, corresponding to the GSV-self, GSV-spatial, and GSV-temporal methods (Fig. 2). Each row shows

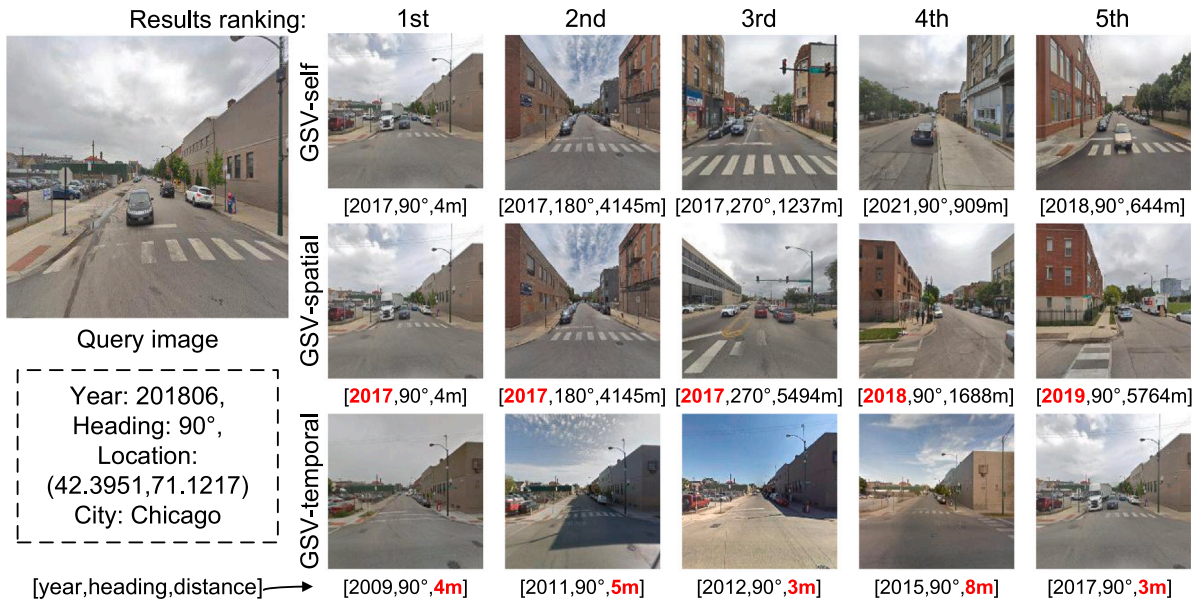


Fig. 4. Comparison of retrieval results using GSV-self, GSV-spatial, and GSV-temporal methods for a given query image (Year: 2018, Heading: 90°, Location: Chicago). Each row corresponds to the top-5 retrieved street view images based on different self-supervised pertained models, ranked by image feature similarity to the query image. The GSV-temporal results are all within a 10-meter radius and have identical heading angles, but correspond to different time periods, demonstrating temporal invariance of the learned image representations. The GSV-spatial results cover a larger geographic area with nearby timeframes, maintaining a consistent overall ambiance.

the top five most similar retrieval results, ranked from left to right (1st to 5th). Below each retrieved image, the year of capture, heading, and actual distance from the query image (in meters) are indicated. The GSV-self method retrieves the nearest street view images based on deep feature similarity. From the comparison, it can be seen that although the retrieved images are from different locations, they are very similar to the query image in feature space, indicating that GSV-self emphasizes overall visual feature similarity without considering consistency in geographic location, heading, or time. The GSV-spatial retrieval results cover a larger geographic area, allowing for greater spatial variation while aiming to maintain a similar overall ambiance and temporal proximity. It can be observed that most of the retrieved street view images are relatively dispersed in space, but the overall ambiance and time are relatively close, reflecting spatial and environmental consistency. This allows GSV-spatial to capture visually similar urban characteristics across different locations. The GSV-temporal retrieval results maintain the same heading and are strictly limited to within a 10-meter radius, highlighting temporal diversity. While the position and heading are mostly unchanged, the retrieved images come from different years. This approach demonstrates sensitivity to temporal changes while keeping other factors consistent, thereby showcasing the variation of the same location across different years.

6.2. What do GSV-temporal and GSV-spatial contrastive objectives learn from GSV?

Our experimental results reveal that different contrastive learning methods excel in different tasks: Temporal contrastive performs exceptionally well in visual place recognition tasks, Spatial contrastive shows better results in macroeconomic prediction tasks, and Self contrastive achieves the best performance in safety perception tasks, confirming our hypothesis that street view images captured at the same location over time enable contrastive learning tasks to uncover the temporal-invariant characteristics of the urban environment. Similarly, spatially proximate street view images from the same period facilitate learning tasks to capture the spatial-invariant neighborhood ambiance, such as the socioeconomic overall ambiance. To further understand how different models allocate their attention to various aspects of the input, we visualized the attention maps in ViT and evaluated the spatial extent of attention using attention distance. This analysis reveals the distinct focus areas of each model, shedding light on their feature extraction preferences.

6.2.1. GSV-temporal learns temporal invariant characteristics, and GSV-spatial learns invariant neighborhood ambiance

To provide a more intuitive and interpretable visualization of what our models learn, we employ Grad-CAM (Selvaraju et al., 2017) to generate attention heatmaps (Fig. 5). This method highlights the image regions most influential to the model's final representation. We selected two street view images from the same location but captured five years apart (2012 vs. 2017), allowing us to observe how each contrastive learning strategy handles temporal changes.

As shown in Fig. 5(a) and (d), the GSV-self model, trained with instance-level contrast, tends to focus on the most visually salient objects in each image independently. For instance, in the 2012 image, its attention is drawn to the dark SUV and prominent building facades. In the 2017 image, its focus shifts to the white van and different storefronts. This indicates that GSV-self learns strong general features but does not inherently distinguish between permanent and transient elements of the scene.

The GSV-temporal model demonstrates a clear ability to learn time-invariant characteristics. In both Fig. 5(b) and (e), the model focuses on permanent structures such as the building architecture on the left and right, the overall street layout, and the horizon. Crucially, it learns to ignore transient objects like cars and pedestrians, which are present in different positions and forms across the years. The attention on vehicles is significantly suppressed compared to GSV-self. This visualization provides evidence that temporal contrast effectively filters out dynamic elements to capture the stable, enduring characteristics of a location.

The GSV-spatial model exhibits a distinctly different pattern. As seen in Fig. 5(c) and (f), its attention is much more holistic and diffuse, spreading across the entire scene. Rather than focusing on specific objects, it captures the overall "atmosphere"—encompassing the buildings, street, sky, and foliage collectively. The attention patterns between 2012 and 2017 are remarkably similar in their broad

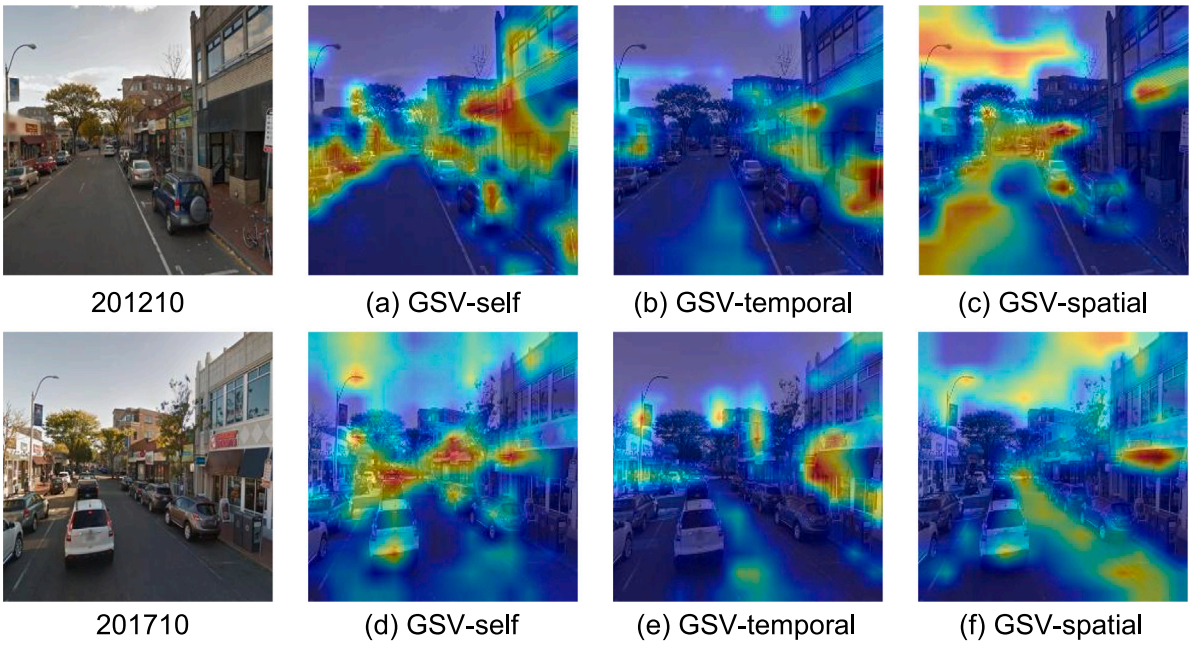


Fig. 5. Grad-CAM visualization of model attention under different contrastive learning strategies. The heatmaps show the focus of three models on street view images of the same location captured at different times (top: 2012, bottom: 2017). (a, d) GSV-self (instance contrast) focuses on salient objects within individual images. (b, e) GSV-temporal (temporal contrast) learns to focus on time-invariant structures, such as building facades, while ignoring dynamic objects like vehicles. (c, f) GSV-spatial (spatial contrast) exhibits a broader, holistic attention, capturing the overall scene ambiance.

scope, suggesting the model learns the invariant spatial context and layout of the neighborhood. This supports our hypothesis that spatial contrast encourages the model to learn the ambient characteristics of an environment rather than focusing on individual, dynamic objects

To provide quantitative support for these visual observations, we evaluate the spatial extent of self-attention using attention distance (Dosovitskiy et al., 2021), which measures the mean distance between query tokens and key tokens, weighted by their respective self-attention scores. This metric helps assess how different contrastive strategies focus on various aspects of the scene. Figs. 6(a) and 6(b) show the attention distances computed for sampled street view images and ImageNet images. Depth corresponds to the network layers in the ViT model, ranging from shallow (Depth 1) to deep layers (Depth 12). Larger attention distances indicate that the model captures more globally distributed features, while smaller distances suggest a focus on local patterns. Specifically, GSV-spatial exhibits the largest attention distance, indicating a tendency to focus on a broader spatial context rather than concentrating on individual objects. In contrast, the attention distances of GSV-temporal and GSV-self decrease sequentially, suggesting a gradual narrowing of focus to capture more specific details within the scenes. Notably, ImageNet-self demonstrates the smallest attention distance, reflecting its pre-training on a dataset primarily consisting of object-centric images, which leads to a greater emphasis on individual objects over the overall spatial arrangement.

6.2.2. GSV-temporal highlights low-frequencies, and GSV-spatial exploits high-frequencies

The low-frequency amplitude of an image represents its large-scale structure and smooth transitions, primarily encompassing the background, gradient regions, and general contours. It reflects the overall form of the image and broad variations in brightness. Low-frequency components are typically key elements in global structure modeling and scene consistency understanding, which is why their amplitude is generally larger. In contrast, the high-frequency amplitude of an image represents finer details, textures, and edges, and is primarily associated with regions of rapid changes in the image, such as boundaries and local contrast variations. Although high-frequency amplitudes are relatively smaller, they are crucial for capturing the sharpness and clarity of the image and often contain noise signals. In this study, we hypothesize that, compared to GSV-spatial, GSV-temporal is more inclined to focus on low-frequency information. This is because temporal-invariant characteristics in street view images rely more on global consistency and invariant structures, while high-frequency information is more susceptible to noise interference in dynamic scenes. To test this hypothesis, we compute the amplitude differences in the Fourier-transformed frequency spectrum of intermediate features across various layers of the ViT backbone, reporting the relative amplitudes of high and low frequencies (Park et al., 2023). Specifically, Figs. 6(c) and 6(d) present the relative amplitude results for ImageNet and GSV images under different contrast strategies.

The results indicate that models pre-trained on ImageNet focus more on high-frequency information, while models pre-trained on GSV emphasize low-frequency information. This difference may stem from the fact that ImageNet images typically center around object categories (e.g., animals, plants, etc.) that require detailed edge and texture detection, thus highlighting high-frequency information. In contrast, street view images feature large-scale street layouts and global structural variations, where the models need to capture more low-frequency information to understand the overall spatial relationships within the scene. Furthermore, we observe that GSV-temporal exhibits the most pronounced sensitivity to low-frequency information. This suggests that the temporal-invariant characteristics prioritize the consistency of static elements, such as street layouts, while being less sensitive to texture variations caused by factors like lighting or seasonality. GSV-self, similar to GSV-temporal, also focuses more on low-frequency information, but due to the need to capture dynamic elements such as pedestrian and vehicular flow, it exhibits a slightly higher relative amplitude compared to GSV-temporal. On the other hand, GSV-spatial shows a stronger focus on high-frequency information. This can be attributed to its lesser sensitivity to the overall street layout, as it is more concerned with capturing consistency in the surrounding environment, which is often conveyed through high-frequency details such as window styles, building facades, and material textures.

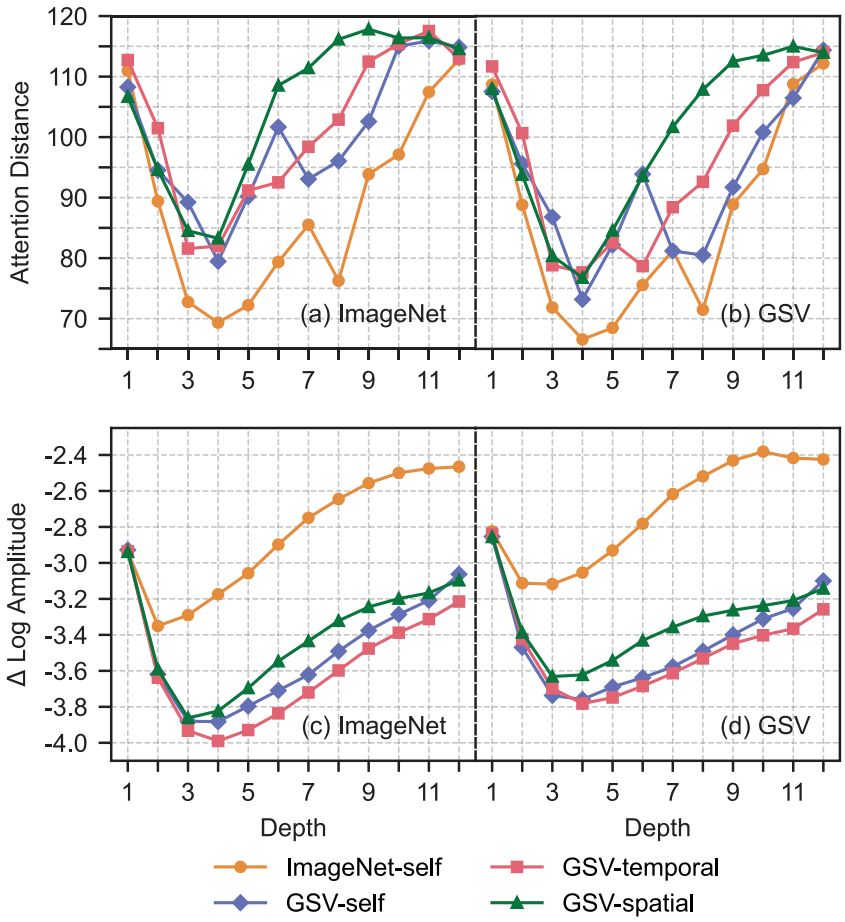


Fig. 6. Visualization of attention distance and Δ Log Amplitude across depths for ImageNet and GSV models. Depth refers to the network layers in the ViT model, from shallow (Depth 1) to deep layers (Depth 12). (a) and (b) display the attention distance, which represents the average spatial range of the attention mechanism in each layer—a larger value indicates that the model attends to more globally distributed features, while smaller values suggest a focus on local patterns. (c) and (d) present the Δ Log Amplitude, where higher values (closer to 0) reflect stronger retention of high-frequency information (e.g., edges, textures), and lower values (more negative) indicate a focus on low-frequency components, representing global structures or smooth transitions.

7. Conclusion

In this work, we proposed a self-supervised learning framework, the spatiotemporal contrastive framework, designed to learn representations from street view imagery. We systematically implemented and evaluated three of its core strategies: Temporal Contrast, Spatial Contrast, and Instance Contrast. Our experimental results demonstrate that these distinct strategies effectively learn features tailored for different urban tasks, achieving significant performance improvements in visual place recognition, socioeconomic prediction, and safety perception. Furthermore, our in-depth analysis provides valuable insights into how each method captures different aspects of the urban environment, emphasizing the importance of targeted learning strategies. This study provides a valuable benchmark for self-supervised learning in urban science and enhances the practical applicability of street view data.

While our implemented strategies perform robustly over typical time scales, we recognize their limitations when considering long-term urban evolution spanning several decades. The Temporal Contrast model, for instance, relies on the assumption that a location's core static features (e.g., buildings) persist over time. This assumption may be challenged in the face of radical urban transformations, such as large-scale demolitions and redevelopment, which can alter a location's visual identity entirely.

This challenge, however, points to a promising future direction that is already conceptualized within our framework. The fourth quadrant of our framework, Spatio-temporal Contrast, is designed precisely to address these long-term dynamics. By learning from different locations at different times, it aims to capture a neighborhood's long-term core identity, a representation more resilient to drastic structural changes. Future work should focus on implementing and evaluating this spatio-temporal strategy. Doing so would extend our framework to model the dynamics of urban change over much longer timescales and unlock new applications in longitudinal urban analysis.

CRediT authorship contribution statement

Yong Li: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Yingjing Huang:** Writing – original draft, Visualization, Resources, Methodology, Data curation. **Fan Zhang:** Writing – review & editing, Writing – original draft, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

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

Acknowledgments

We also acknowledge the financial support from the National Natural Science Foundation of China (Grant No. 42371468). This work was supported by the High-performance Computing Platform of Peking University.

Data availability

Data will be made available on request.

References

Arandjelović, R., Gronat, P., Torii, A., Pajdla, T., & Sivic, J. (2018). NetVLAD: CNN architecture for weakly supervised place recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1437–1451. http://dx.doi.org/10.1109/TPAMI.2017.2711011.

Ayush, K., Uzkent, B., Meng, C., Tanmay, K., Burke, M., Lobell, D., & Ermon, S. (2021). Geography-aware self-supervised learning. In *2021 IEEE/CVF international conference on computer vision* (pp. 10161–10170). Montreal, QC, Canada: IEEE, http://dx.doi.org/10.1109/ICCV48922.2021.01002.

Boeing, G. (2017). OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*, *65*, 126–139.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9650–9660).

Chacra, D. A., & Zelek, J. (2018). Municipal infrastructure anomaly and defect detection. In *2018 26th European signal processing conference* (pp. 2125–2129). Rome: IEEE.

Chen, X., & He, K. (2021). Exploring simple siamese representation learning. In *2021 IEEE/CVF conference on computer vision and pattern recognition* (pp. 15745–15753). Nashville, TN, USA: IEEE, http://dx.doi.org/10.1109/CVPR46437.2021.01549.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In H. D. III, & A. Singh (Eds.), *Proceedings of machine learning research: Vol. 119*, *Proceedings of the 37th international conference on machine learning* (pp. 1597–1607). PMLR.

Chen, Z., Liu, L., Sa, I., Ge, Z., & Chli, M. (2018). Learning context flexible attention model for long-term visual place recognition. *IEEE Robotics and Automation Letters*, *3*(4), 4015–4022. http://dx.doi.org/10.1109/LRA.2018.2859916.

Chen, X., Xie, S., & He, K. (2021). An empirical study of training self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision* (pp. 9620–9629). http://dx.doi.org/10.1109/ICCV48922.2021.00950.

Cheng, J., Tsai, Y.-H., Wang, S., & Yang, M.-H. (2017). Segflow: Joint learning for video object segmentation and optical flow. In *Proceedings of the IEEE international conference on computer vision* (pp. 686–695).

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3213–3223).

Deng, J., Dong, W., Socher, R., Li, L. J., Kai Li, & Li Fei-Fei (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). Miami, FL: IEEE.

Deuser, F., Habel, K., & Oswald, N. (2023). Sample4Geo: Hard negative sampling for cross-view geo-localisation. In *2023 IEEE/CVF international conference on computer vision* (pp. 16801–16810). Paris, France: IEEE, http://dx.doi.org/10.1109/ICCV51070.2023.01545.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International conference on learning representations*. URL https://openreview.net/forum?id=YicbFdNTTy, p. ..

Dubey, A., Naik, N., Parikh, D., Raskar, R., & Hidalgo, C. A. (2016). Deep learning the city: Quantifying urban perception at a global scale. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer vision – ECCV 2016* (pp. 196–212). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-319-46448-0_12.

Fan, Z., Zhang, F., Loo, B. P. Y., & Ratti, C. (2023). Urban visual intelligence: Uncovering hidden city profiles with street view images. *Proceedings of the National Academy of Sciences*, *120*(27), Article e2220417120.

Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., & Fei-Fei, L. (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), 13108–13113. http://dx.doi.org/10.1073/pnas.1700035114.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., & He, K. (2018). Accurate, large minibatch SGD: Training ImageNet in 1 hour. http://dx.doi.org/10.48550/arXiv.1706.02677, arXiv:1706.02677.

Grill, J. B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. (2020). Bootstrap your own latent-a new approach to self-supervised learning. In *Advances in neural information processing systems: vol. 33*, (pp. 21271–21284).

Guo, D., Yu, Y., Ge, S., Gao, S., Mai, G., & Chen, H. (2024). SpatialScene2Vec: A self-supervised contrastive representation learning method for spatial scene similarity evaluation. *International Journal of Applied Earth Observation and Geoinformation*, *128*, Article 103743.

He, K., Chen, X., Xie, S., Li, Y., Dollar, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition* (pp. 15979–15988). New Orleans, LA, USA: IEEE, http://dx.doi.org/10.1109/CVPR52688.2022.01553.

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF conference on computer vision and pattern recognition* (pp. 9726–9735). Seattle, WA, USA: IEEE, http://dx.doi.org/10.1109/CVPR42600.2020.00975.

Huang, Y., Wen, Z., Chi, Y., & Liang, Y. (2024). How transformers learn diverse attention correlations in masked vision pretraining. In *ICML 2024 workshop on theoretical foundations of foundation models*.

Huang, Y., Zhang, F., Gao, Y., Tu, W., Duarte, F., Ratti, C., Guo, D., & Liu, Y. (2023). Comprehensive urban space representation with varying numbers of street-level images. *Computers, Environment and Urban Systems*, *106*, Article 102043.

Klemmer, K., Rolf, E., Robinson, C., Mackey, L., & Rußwurm, M. (2024). SatCLIP: Global, general-purpose location embeddings with satellite imagery. arXiv:2311.17179.

Liu, Y., Zhang, X., Ding, J., Xi, Y., & Li, Y. (2023). Knowledge-infused contrastive learning for urban imagery-based socioeconomic prediction. arXiv:2302.13094.

Loshchilov, I., & Hutter, F. (2017). SGDR: Stochastic gradient descent with warm restarts. In *International conference on learning representations*. p. ..

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *International conference on learning representations*. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Lowry, S., Sünderhauf, N., Newman, P., Leonard, J. J., Cox, D., Corke, P., & Milford, M. J. (2015). Visual place recognition: A survey. *IEEE Transactions on Robotics*, *32*(1), 1–19.

Mai, G., Lao, N., He, Y., Song, J., & Ermon, S. (2023). Csp: Self-supervised contrastive spatial pre-training for geospatial-visual representations. In *International conference on machine learning* (pp. 23498–23515). PMLR.

Manas, O., Lacoste, A., Giro-i Nieto, X., Vazquez, D., & Rodriguez, P. (2021). Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *2021 IEEE/CVF international conference on computer vision* (pp. 9394–9403). Montreal, QC, Canada: IEEE, http://dx.doi.org/10.1109/ICCV48922.2021.00928.

Mans Larsson, M., Stenborg, E., Hammarstrand, L., Pollefeys, M., Sattler, T., & Kahl, F. (2019). A cross-season correspondence dataset for robust semantic segmentation. In *2019 IEEE/CVF conference on computer vision and pattern recognition* (pp. 9524–9534). Long Beach, CA, USA: IEEE, http://dx.doi.org/10.1109/CVPR.2019.00976.

Naik, N., Kominers, S. D., Raskar, R., Glaeser, E. L., & Hidalgo, C. A. (2017). Computer vision uncovers predictors of physical urban change. *Proceedings of the National Academy of Sciences*, *114*(29), 7571–7576. http://dx.doi.org/10.1073/pnas.1619003114.

van den Oord, A., Li, Y., & Vinyals, O. (2019). Representation learning with contrastive predictive coding. arXiv:1807.03748.

Park, N., Kim, W., Heo, B., Kim, T., & Yun, S. (2023). What do self-supervised vision transformers learn? In *The eleventh international conference on learning representations*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763). PMLR.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618–626).

Stalder, S., Volpi, M., Büttner, N., Law, S., Harttgen, K., & Suel, E. (2024). Self-supervised learning unveils urban change from street-level images. *Computers, Environment and Urban Systems*, *112*, Article 102156.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *58*(1), 267–288.

Wang, X., Jabri, A., & Efros, A. A. (2019). Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2566–2576).

Wang, Z., Li, H., & Rajagopal, R. (2020). Urban2vec: Incorporating street view imagery and pois for multi-modal urban neighborhood embedding. In *Proceedings of the AAAI conference on artificial intelligence: vol. 34*, (pp. 1013–1020).

Wang, Y., Zhang, J., Kan, M., Shan, S., & Chen, X. (2020). Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12275–12284).

Warburg, F., Hauberg, S., López-Antequera, M., Gargallo, P., Kuang, Y., & Civera, J. (2020). Mapillary street-level sequences: A dataset for lifelong place recognition. In *2020 IEEE/CVF conference on computer vision and pattern recognition* (pp. 2623–2632). http://dx.doi.org/10.1109/CVPR42600.2020.00270.

Wu, Z., Xiong, Y., Yu, S. X., & Lin, D. (2018). Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3733–3742).

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., & Hu, H. (2022). Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9653–9663).

Zaffar, M., Ehsan, S., Milford, M., & McDonald-Maier, K. D. (2021). Memorable maps: A framework for re-defining places in visual place recognition. *IEEE Transactions on Intelligent Transportation Systems*, *22*(12), 7355–7369. http://dx.doi.org/10.1109/TITS.2020.3001228.

Zhang, Y., Li, Y., & Zhang, F. (2024). Multi-level urban street representation with street-view imagery and hybrid semantic graph. *ISPRS Journal of Photogrammetry and Remote Sensing*, *218*, 19–32.

Zhang, F., Salazar-Miranda, A., Duarte, F., Vale, L., Hack, G., Chen, M., Liu, Y., Batty, M., & Ratti, C. (2024). Urban visual intelligence: Studying cities with artificial intelligence and street-level imagery. *Annals of the American Association of Geographers*, *114*(5), 876–897.

Zhang, F., Zhou, B., Liu, L., Liu, Y., Fung, H. H., Lin, H., & Ratti, C. (2018). Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning*, *180*, 148–160.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.