CityPulse: Fine-Grained Assessment of Urban Change with Street View Time Series

Tianyuan Huang[1*], **Zejia Wu**[2*], **Jiajun Wu**[1], **Jackelyn Hwang**[1], **Ram Rajagopal**[1]

[1]Stanford University [2]University of California San Diego

{tianyuah, jihwang, ramr}@stanford.edu, zew024@ucsd.edu, jiajunwu@cs.stanford.edu

Abstract

Urban transformations have profound societal impact on both individuals and communities at large. Accurately assessing these shifts is essential for understanding their underlying causes and ensuring sustainable urban planning. Traditional measurements often encounter constraints in spatial and temporal granularity, failing to capture real-time physical changes. While street view imagery, capturing the heartbeat of urban spaces from a pedestrian point of view, can add as a high-definition, up-to-date, and on-the-ground visual proxy of urban change. We curate the largest street view time series dataset to date, and propose an end-to-end change detection model to effectively capture physical alterations in the built environment at scale. We demonstrate the effectiveness of our proposed method by benchmark comparisons with previous literature and implementing it at the city-wide level. Our approach has the potential to supplement existing dataset and serve as a fine-grained and accurate assessment of urban change.

Introduction

Our cities are evolving, and understanding how cities change at a granular level has far-reaching societal impact — from facilitating better urban planning and infrastructure assessment to enabling more sustainable social and environmental interventions (**??**). Current measurements of urban change rely on datasets ranging from survey data such as American Community Survey (ACS), to government open data like construction permits, to remote sensing data such as satellite and aerial imagery. However, survey data often fall short of spatial and temporal granularity (**?**), and top-down perspectives from the remote sensing data may not adequately represent the street-level changes that directly impact the daily lives of urban residents. And some construction permits data are not universally accessible. Street view imagery, on the other hand, offers a high-resolution and frequently updated visual representation of urban environments from a ground-level perspective (**?**). By curating and analyzing the time series data of street view imagery, we can establish a more precise and direct proxy for how cities evolve over time.

Previous studies on street view change detection have primarily focused on comparing pairwise images from identical locations but at different times, utilizing pixel-level annotations (**??**), which is similar to change detection tasks using satellite imagery (**??**). Recent works have also demonstrated the applicability of street view pairwise change detection by collecting large-scale historical street view datasets and apply them on a range of urban applications, such as mapping out physical improvements and declines in cities, as well as correlating with socio-economic attributes and neighborhood gentrification status (**??**).

However, unlike satellite imagery, street view imagery can be more susceptible to noisy signals, such as varying camera angles and noisy background elements like shadows and lighting changes. Additionally, existing street view datasets for change detection tasks are often limited in both spatial and temporal scales due to the fact that pixel-level semantic annotations can be costly. Such constraints hamper the model's generalizability and scalability, making it chal-

Figure 1: Detection of urban change points using street view time series. Red bounding boxes highlight transformations in the built environment at each location. By aggregating these detected change points within a neighborhood, we can evaluate the temporal dynamics of urban development.

*These authors contributed equally.

lenging to directly apply them to downstream tasks and thus restricting their potential social impacts.

To address these challenges, we first introduce a comprehensive multi-city street view time series dataset with image-level semantic labels, and then propose an end-to-end framework to detect urban change with street view data. We demonstrate the effectiveness of our approach with a fine-grained assessment of urban change across Seattle, Washington. Specifically, our major contributions in this study are threefold:

- We collect and curate a Google Street View (GSV) time series dataset, covering more than 1000 coordinates across 6 different cities, which is the largest street view change detection dataset available up to date. Each street view time series is labeled with change or no change on the image level, and each series has an average length of 10 images, covering a time interval of 16 years (from 2007 to 2023). We further validate the benefits of the time series data over pairwise data in our experiment.
- We propose an end-to-end change detection pipeline that effectively learns feature representations with semantic contexts from street view time series data, which allows the model to not only extract object shape, color, and structural information of the built environment, but also mitigate noisy effects from lighting changes and angle misalignment, enhancing the overall robustness of the change detection.
- Our method enables scalable applications for urban scene change detection, providing a more accurate proxy for assessing neighborhood socio-economic status changes. We demonstrate the efficacy of our approach by evaluating its correlation with social-demographic data and comparing against construction permits through a case study in Seattle.

Related work

Urban change assessment

Measuring physical change in urban environments offers profound insights into urban policies and economics, illuminating housing value trends, shifts in urban areas' roles, and spatial segregation effects (**??**). It also has significant values in a various downstream tasks, such as detecting neighborhood gentrification (**?**), monitoring the disaster recovery (**?**). Prior research primarily utilizes satellite imagery for large-scale urban change detection (**??**). However, satellite data are hindered by constraints in spatial and temporal resolution, and lack detail on fine-grained and street-level changes. Several researchers use building permits data as a fine-grained proxy for physical urban change (**??**). While these data also have limitations in availability and spatial coverage and may not accurately represent actual changes due to potential delays, as indicated by our evaluations.

Street view imagery

Street view imagery have been used in a wide range of applications in urban studies, such as quantifying urban greenery (**?**), indicating region functions (**?**), revealing economic and social-demographic patterns(**???**), predicting populace' well-beings (**?**) and estimating building energy efficiency (**?**). It demonstrates substantial value as a medium closely reflecting human perception of the city. Moreover, recent studies analyzed historical street view data in the temporal dimension to map physical improvement and decline in the built environment and uncover how cities changed over time (**???**). However, existing methods rely on comparing pairs of historical street views for each location rather than a comprehensive time series of street view data, which is limited in tracking the complete range of transformations within urban environments.

Change detection

Change detection is commonly presented in the field of remote sensing, as the task to identify changes between the pixel-level features of two temporally separated images. Previous works have trained convolutional network, recurrent network and Siamese network for change detection task on satellite imagery (**???**). Recent works also explored self-supervised pretraining and unsupervised learning in change detection to rely less on labels and generate meaningful representations for other downstream tasks (**??**). Unlike satellite imagery, Street view change detection faces additional challenges such as noisy signals including shades and angle misalignments due to the less fixed and more variable nature of acquiring street-level visual data. Previous works introduced benchmarking datasets for scene change detection and adopted deconvolutional networks or temporal attention networks (**???**). However, current research not only lacks comprehensive benchmark datasets with expansive spatial and temporal coverage, but also falls short of the model generalizability, limiting their societal impact. To tackle this, we introduce the largest image-level change detection dataset to date, featuring a complete time series of street views for each sampled location, applied at the city scale.

Methods

Problem Statement

Definition 1 ***Street view time series.*** *Each street view time series, comprises n street view images depicting the consistent street-level scene $s^{(i)} = (s_1^{(i)}, s_2^{(i)}, ..., s_n^{(i)})$. These images are chronologically arranged such that $s_k^{(i)}$ corresponds to the timestamp $t_k^{(i)}$.*

Definition 2 ***Urban change point.*** *In the street view time series $s^{(i)}$, the image $s_c^{(i)}$ is identified as an urban change point if the built environment in $s_c^{(i)}$ exhibits deviations (e.g., building constructions) relative to preceding images.*

Our objective is to accurately and efficiently detect urban change points in street view time series. To achieve this, we begin by creating a large street view time series dataset, followed by proposing an end-to-end training and evaluation pipeline.

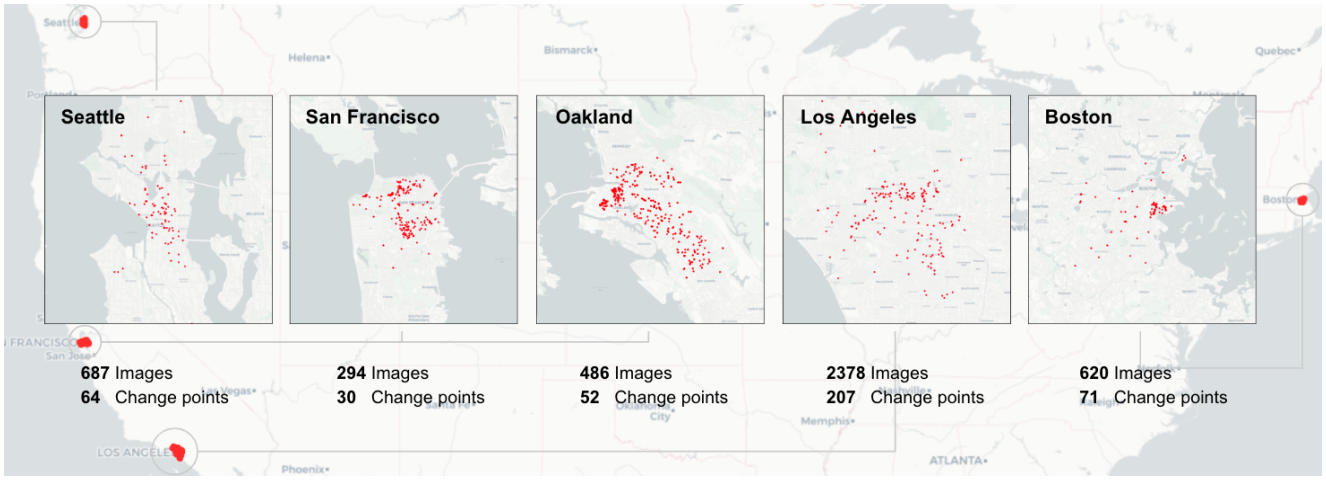

Figure 2: Geo-spatial distribution of our street view time series dataset across 5 different cities in the US. Locations are selected based on open-access building footprint data, and historical Google Street View imagery from these coordinates is comprehensively downloaded and labeled with urban change points.

Street view time series dataset

To start, we sample the geospatial coordinate for each street-level scene $s^{(i)}$. Specifically, the coordinate of each building is determined by computing the centroid of its footprint polygon, using the Microsoft building footprint dataset. After locating scene $s^{(i)}$, we gather all the available historical street view metadata and subsequently download the associated images using their panoid ID. For each scene $s^{(i)}$, we retrieve the nearest-photographed panorama. The image heading is subsequently determined, facing the building from the panorama's coordinate. All our street view images and meta data are sourced from the Google Static Street View API.

Dataset	# images	# pairs	Areas	Timeframe
TSUNAMI	200	100	2 cities	<1 year
PSCD	1540	770	1 city	<10 years
Ours	**4465**	**25423**	**5 cities**	**16 years**

Table 1: Scene change detection dataset comparison

In total, we select 931 locations and retrieve their corresponding street view time series, which consist of $10,878$ images. We then annotate each time series $s^{(i)}$ to identify the urban change points $s_c^{(i)}$. Among them, 371 time series have been labeled with a total of 433 urban change points, while the remaining ones exhibit no substantial urban change. Figure 2 demonstrates the geo-spatial distribution of our sampled street views. As is shown in table 1, our dataset not only consists of more image pairs compared with previous scene change detection datasets TSUNAMI (**?**) and PSCD (**?**), but also covers a significantly broader spatial and temporal scope.

Data partitioning. To train our change detection model on street view time series dataset labeled with urban change points, we introduce a partitioning scheme for generating training and evaluation sets as is shown in Figure 3. For every street view time series $s^{(i)}$, we segment the views from $s_1^{(i)}$ to $s_n^{(i)}$ based on their occurrence relative to $s_c^{(i)}$. More precisely, suppose the time series $s^{(i)}$ has q urban change points denoted as $s_{c_1}^{(i)}, \ldots, s_{c_q}^{(i)}$, we allocate the street view $s_j^{(i)}$ $(1 \le j \le n)$ to segment $seg(s_j^{(i)})$ as follows:

$$seg(s_j^{(i)}) = \begin{cases} 1 & \text{if } j < c_1 \\ k & \text{if } c_1 \le j < c_q \text{ and } c_{k-1} \le k < c_k \\ q+1 & \text{if } j \ge c_q \end{cases} \tag{1}$$

For each street view time series $s^{(i)}$, we then generate a set of pairwise street view pairs by considering all combinations from $s_1^{(i)}$ to $s_n^{(i)}$. Each pair of samples is sorted in chronological order based on their timestamps, resulting in pairs like $(s_1^{(i)}, s_2^{(i)})$. The total number of such combinations for the time series $s^{(i)}$ with n street views is $\binom{n}{2}$. The labeling of these pairwise samples is determined by their associated segments as follows:

$$\text{LABEL}(s_a^{(i)}, s_b^{(i)}) = \begin{cases} 1 & \text{if } seg(s_a^{(i)}) \neq seg(s_b^{(i)}) \\ 0 & \text{if } seg(s_a^{(i)}) = seg(s_b^{(i)}) \end{cases} \tag{2}$$

Change detection model

To classify each pairwise pair $(s_a^{(i)}, s_b^{(i)})$, we adopt a Siamese network to include a twin DINOv2 (**?**) backboned module to realize a non-linear embedding from the input domain and a final linear layer transforming the concatenation of both images' hidden vectors and their distance, represented by their element-wise difference, into a scalar predictor as follows:

$$\mathbf{h}_L^{(i)} = \left[(\mathbf{h}_{l,L}^{(i)})^\top, (\mathbf{h}_{e,L}^{(i)})^\top, (\mathbf{h}_{l,L}^{(i)} - \mathbf{h}_{e,L}^{(i)})^\top \right]^\top \tag{3}$$

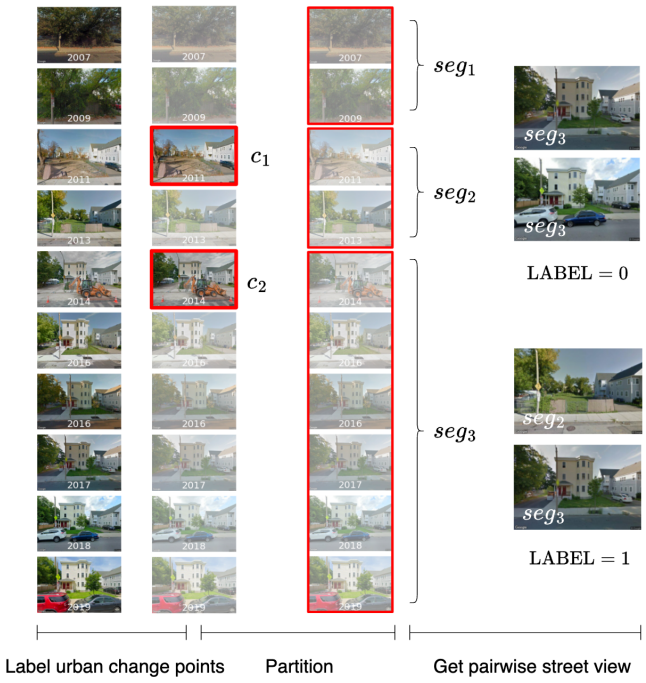

Figure 3: Partitioning of street view time series data. All possible pairwise combinations of street view samples are generated from each time series. Each pair's label is assigned based on its corresponding position with the urban change points.

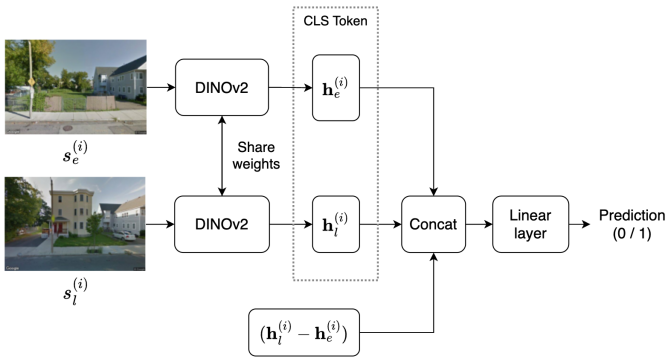

Figure 4: Overview of the change detection model architecture. Pairs of input images are processed using Siamese-based networks with DINOv2 as the backbone. The CLS tokens serve as the image representation, with a subsequent linear layer projecting them to a prediction score.

Figure 4 visualizes the model architecture. We adopt a cross-entropy loss function to train such a urban change classifier, and let $\text{LABEL}(s_a^{(i)}, s_b^{(i)})$ be the label for the street view pair $(s_a^{(i)}, s_b^{(i)})$.

Experiments

To evaluate the efficacy of our proposed method, we conduct experiments from three perspectives: 1) Backbone models — we benchmark the performance of selected visual foundational models in the context of street view change detection tasks. 2) Street view time series data — we employ experiments to substantiate the advantage of time series data as a natural form of data augmentation (**?**), compared with results achieved through artificial data augmentation. 3) Self-supervised pre-training — we explore 2 pre-train methodologies using a larger-scale unlabeled street view dataset in order to evaluate the performance of a domain-specific pre-trained models for our change detection task.

Training details

Street view images differ from satellite imagery and high-quality object images in that they often have a lower signal-to-noise ratio, primarily due to varying camera positions and environment conditions as shown earlier. As a result, evaluating on a small-scale test set could suffer from a significant variance. To ensure a robust assessment, we randomly select 50% of the street view time series in our dataset as the test set. It includes 25 locations in Seattle, 13 locations in San Francisco, 21 locations in Oakland, 97 locations in Los Angeles, and 29 locations in Boston, constituting a total of $12,221$ image pairs. The remaining data are allocated with 90% as the training set for model fine-tuning and 10% as the validation set. During fine-tuning, we employ the Adam optimizer to train models with a learning rate set at 1×10^{-5} and a batch size of 16. The global norm of gradients is clipped to be ≤ 0.5, and we use random weight averaging for optimization. Our training and evaluation are conducted on 4 Nvidia Tesla T4 GPUs. For all the backbone models, we experimented with two common approaches: global fine-tuning and linear probing, i.e. training with the backbone network frozen.

Backbone models. We initiate our evaluation by assessing the performance of 4 pre-trained generic visual models—ResNet101 (**?**), DINO (**?**), CLIP (**?**), and DINOv2 (**?**)—as backbone networks. For ViT-based models such as DINO and DINOv2, we experiment using the CLS Token as the backbone output.

Time series data. To validate the advantages of our proposed street view time series dataset, we constructed a pair-based dataset similar to TSUNAMI (**?**) and PSCD (**?**) dataset. Specifically, for each street view time series, we randomly sample 2 images as a pair. We conduct model fine-tuning on this pair-based dataset and evaluated its performance on the test set described earlier. To align the pair-based dataset with the size of our time series dataset, we randomly apply a combination of standard image augmentation techniques, including horizontal flip, color jitter, grayscale, and Gaussian blur. It seeks to validate our hypothesis that time series images, serving as natural augmentation, are more effective than artificial augmentations to supervise change detection model amidst noisy signals, thus bolstering its robustness.

Self-supervised pre-training. Recent studies on self-supervised pre-training highlight its efficacy to extract image features when labels are limited and enhance performance

Models	Linear Probing				Fine-Tuning			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
ResNet101	69.18	62.43	**92.17**	74.44	84.52	79.92	**91.07**	85.13
DINO (ViT-B/16)	82.33	82.24	81.25	81.74	86.24	**93.30**	77.28	84.54
CLIP	86.01	86.95	83.85	85.37	87.06	91.52	80.91	85.89
DINOv2 (ViT-B/14)	**88.20**	**92.08**	82.88	**87.24**	**88.85**	92.77	83.62	**87.96**

Table 2: Performance of different backbone models using linear probing and fine-tuning.

Data	Data Augmentation	# pairs	Accuracy	Precision	Recall	F1-Score
Pairwise data	None	336	85.99	85.28	**86.07**	85.67
Pairwise data	HorizontalFlip + ColorJitter + GrayScale + GaussianBlur	11922	85.63	88.60	80.89	84.57
Time series data	None	11922	**88.85**	**92.77**	83.63	**87.96**

Table 3: Street view time series vs. pairwise data.

in downstream tasks. Specifically, 2 primary branches of self-supervised pre-training are pursued: intra-image self-supervised training (**??**) and discriminative self-supervised learning (**?**). Correspondingly, we adapt 2 pre-training procedures on street view data and benchmark their performance on the change detection task — StreetMAE and StreetBYOL. StreetMAE uses masked autoencoders (**?**) to reconstruct randomly masked patches in street view imagery. It also incorporates temporal encoding to represent each street view time series as a contextual sequence. StreetBYOL, on the other hand, is a self-distillation approach building upon the online and target networks (**?**). While retaining pivotal components such as the prediction head and the stop gradient mechanism, we try add an unsupervised segmentation head (**?**) to identify building pixels and feed them alongside the original images into the networks in experiment seg+StreetBYOL. We adopt the ViT-B/16 architecture as the backbone network and initialize it with the parameters from DINO pre-trained model. Pre-training is conducted on an unlabeled dataset comprising 150,000 street view images randomly sampled in our studied areas. To mitigate noise interference, we apply a filtering process to remove images where the proportion of building pixels was less than 2%. After the pre-training phase, we plug it into the Siamese network and fine-tune the model on our labeled training set.

Pre-training	Accuracy	Precision	Recall	F1-Score
DINOv2 (ViT-B/14)	**88.85**	**92.77**	**83.62**	**87.96**
StreetMAE	78.49	81.97	71.54	76.40
StreetBYOL	86.25	91.98	78.62	84.78
Seg+StreetBYOL	87.42	91.03	82.27	86.43

Table 4: Performance of different pre-train methods.

Results and discussion

As shown in Table 2, DINOv2 has demonstrated the best performance in our evaluation, achieving 88.85% accuracy through fine-tuning. Notably, considering the presence of challenges like shadows and occlusions in the images, human performance in this change detection task is approximately around 90% during our labeling process. This observation suggests that fine-tuning DINOv2 as the backbone network has enabled the model to approach human-level performance. Furthermore, freezing the DINOv2 network and training only the linear layers surpass the fine-tuning outcomes of all other backbone networks, strongly affirming the capacity of DINOv2 to generate potent visual features suitable for change detection.

Figure 5: Sampled prediction results. Our proposed change detection model effectively identifies structural changes in buildings, while filtering our random variations such like lighting, shadows, vegetation, and vehicles.

We find the performance of the change detection model fine-tuned on the pairwise dataset is significantly lower than its performance attained after fine-tuning on our time series dataset, as illustrated in table 3. Moreover, augmenting the dataset using artificial techniques can lead to adverse effects. As a form of natural data augmentation, time series data equips the model with sufficient information to identify and filter out irrelevant variations that occur over time, such as changes in lighting, vegetation, and vehicles as is shown in Figure 5, which guides the model to focus on more temporally stable elements such as building structures. These results validate the critical role of our proposed time series data in the context of street view image change detection task.

The performance of the street-view pre-trained models is presented in Table 4. The results of StreetMAE are signif-

Figure 6: Assessing urban change in Seattle. **Left:** Location of approximately 800k sampled street view images, each represented by a blue dot. **Middle:** Results from deploying our change detection model on the sampled images to pinpoint urban changes shown in red bounding boxes. **Right:** Change points, aggregated at the census tract level, with color denoting the proportion of street view time series that have been identified as change.

icantly lower compared to those of StreetBYOL. This may be because patch reconstruction process is more prone to learning color and texture information, which aligns with the noise we aim to eliminate in change detection, rather than building structure. The addition of the semantic segmentation module leads to a performance enhancement in StreetBYOL. Nevertheless, domain-specific pre-trained models, whether StreetMAE or StreetBYOL, do not surpass the performance of the generic visual model DINOv2. This can be attributed to the smaller training data size used for domain-specific pre-training compared to generic visual models. Specifically, the inherently noisy street view images, with their complex and cluttered scenes, make it difficult for models to grasp fundamental concepts like shape, location, and architecture from limited data.

Case study: Assessing urban change in Seattle

To evaluate the generalizability of our proposed change detection model on a large scale, we prepare a large-scale street view time series data for the city of Seattle, Washington. Figure 6 demonstrates the sampling process. We then apply our change detection model to identify urban change points: In total, we detect $11,838$ change points from $795,919$ sampled images in Seattle.

Construction permits data. As previously noted, previous works frequently rely on construction permit data as a detailed proxy for urban evolution. To further compare them with the results from our proposed change detection model, we obtain construction permit data from the Seattle city government's online permit center. Each entry in the permit data provides details such as the date of issuance, permit category, geospatial coordinates, estimated cost, and other requisite information as mandated by the government. As a data pre-processing step, we keep the "new", "alteration" and "addition" categories to align with the definition of urban change points. Additionally, we curated a subset of permits that had a total estimated cost exceeding $\$100,000$ within a single year. This approach allows us to compare our findings with both the complete permit dataset and the high-value permits, the latter of which are more probable to signify visible physical alterations.

Correlation with social-demographic data. We prepare social-demographic data at the census tract level from the American Community Survey (ACS) 5-year estimates. Specifically, we select population size and median household incomes as our target variables, and calculate their relative percentage change from 2009 to 2021 for each census tract in Seattle. To quantify the linear correlation between proxies for urban change and shifts in socio-demographics, we compare three distinct proxies: the entire set of permits, high-value permits (those exceeding \$100k), and the percentage of locations with urban change points identified using our proposed methodology. As is shown in Figure 7, both the entire set of permits and the high-value permits fail to show a linear correlation with the change of median household income and population size in each census tract with p value larger than 0.05 and R^2 close to 0. While the change points results from our proposed method reach an R^2 of 0.19 and 0.15 for median household income change and population size change respectively, and achieve a p value much less than 0.05 supporting the statistical significance. These results not only indicate that the detected urban

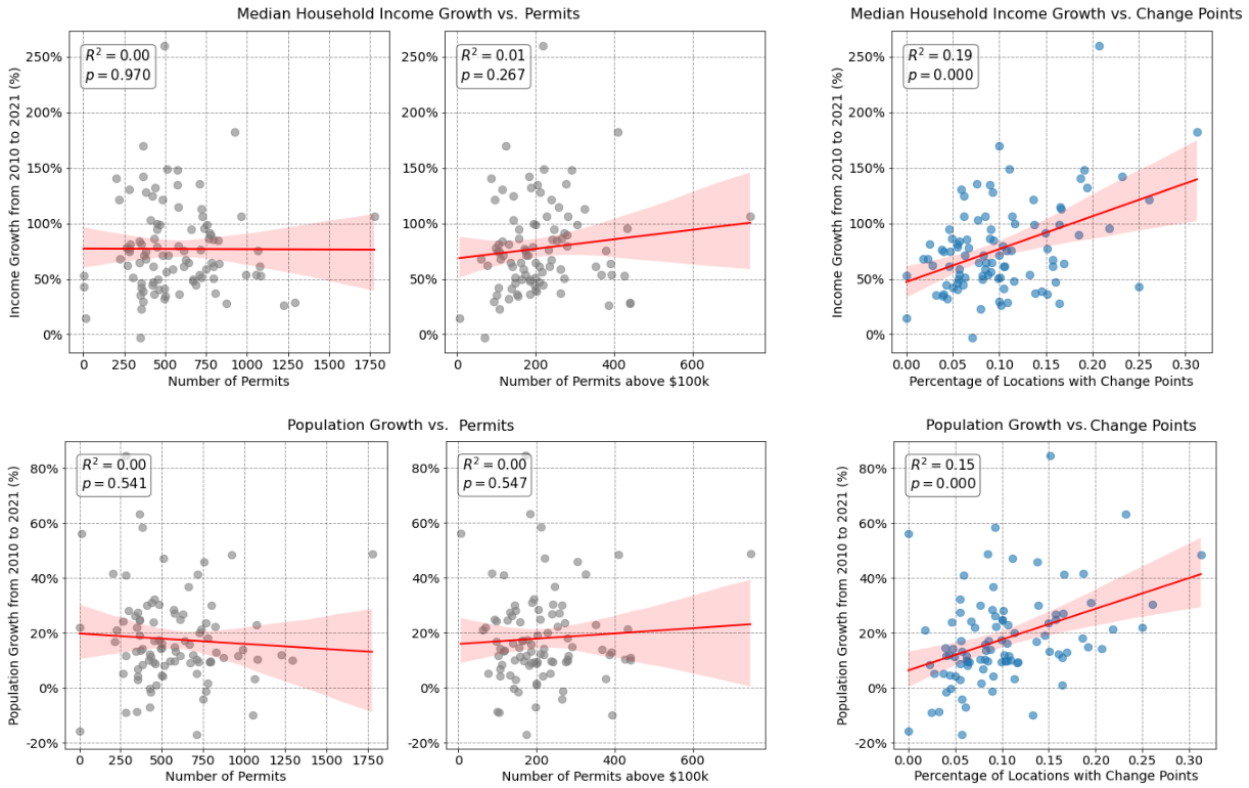

Figure 7: Linear correlation with socio-demographic indicators. **Top:** Median household income. **Bottom:** Population size. Each dot represents a Seattle census tract. The change detection results show statistically significant correlations with socio-demographic metrics, in contrast to construction permit data which lacks such correlation.

change points provide a more accurate assessment of real-world urban transformations and socio-economic shifts, but also validates that our proposed change detection model can effectively complement existing proxies as a credible indicator of urban change.

Conclusion and Future Work

In this work, we propose a framework to assess fine-grained urban change at scale with street view time series. We have curated the largest street-level scene change detection dataset by far, and proposed an end-to-end change detection pipeline to identify urban change points at scale. We validate the proposed model by correlating with social-demographic data and prove its potential as a high-definition, up-to-date, and on-the-ground visual proxy of urban change.

While our data-driven approach provides a novel method to assess urban change, it is still subjective to a few limitations: 1) Street view data focus on changes observable at the street level, excluding alterations that might be non-visible, such as interior renovations. 2) The spatial-temporal distribution of Google Street View data is not consistent. Since its debut in 2007, Google has frequently updated its imagery in countries such as the US, but has been less consistent in updating images in many other regions, especially in developing countries. Despite these limitations, we believe our proposed method offer a comprehensive and extensive resource for urban change detection task, helping expand its social impact and advance sustainable development goals. In future works, we can explore multi-task models to identify changes in a wider array of objects, enhancing the applicability to a broader range of downstream tasks in cities.

Acknowledgments

This project was supported by the Google Cloud Grant from the Stanford Institute for Human-Centered Artificial Intelligence. The author would like to thank Zhecheng Wang, Sarthak Kanodia and Timothy Dai for their extensive guidance. Blanditiis magni perferendis minus debitis deserunt explicabo impedit ipsum iste aliquid rerum, unde suscipit nesciunt libero modi sint rerum.Vel tempora earum veniam debitis perspiciatis, voluptas et sunt delectus vitae aperiam reprehenderit.Quisquam eveniet numquam repudiandae mollitia voluptatibus incidunt saepe pariatur labore, dolores fugit voluptatum recusandae optio velit assumenda esse ipsam atque consectetur doloremque, cum impedit molestias mollitia atque earum repudiandae harum unde ea sint.Esse voluptatem aliquam quisquam, cum quasi illum, hic eligendi totam consectetur suscipit placeat eius ullam dolore sint tempore amet?Minima quam aliquid consequatur ad itaque neque nobis, voluptatum cupiditate eos placeat quaerat in recusandae consequuntur, suscipit adipisci quis dolores incidunt nisi et facilis molestias laboriosam, volup-

tate porro cumque.Beatae deserunt facilis voluptas laudantium obcaecati nesciunt corrupti neque esse repellendus, facilis veniam a deserunt dolores corporis.