Discovering Underground Maps from Fashion

Utkarsh Mall[1, 2] Kavita Bala[1] Tamara Berg[3] Kristen Grauman[2,4]
[1]Cornell University, [2]Facebook AI Research, [3]Facebook, [4] University of Texas at Austin
utkarshm@cs.cornell.edu kb@cs.cornell.edu tlberg@fb.com grauman@cs.utexas.edu

Abstract

The fashion sense—meaning the clothing styles people wear—in a geographical region can reveal information about that region. For example, it can reflect the kind of activities people do there, or the type of crowds that frequently visit the region (e.g., tourist hot spot, student neighborhood, business center). We propose a method to create underground neighborhood maps *of cities by analyzing how people dress. Using publicly available images from across a city, our method automatically segments the map into neighborhoods with a similar fashion sense. Our approach further allows discovering insights about a city, such as detecting distinct neighborhoods (what is the most unique region of NYC?) and answering analogy questions between cities (what is the "Downtown LA" of Bogota?). We also present two new underground map benchmarks derived from non-image data for 37 cities worldwide. Our method shows promising results on both these benchmarks as well as experiments with human judges.*

"The map is not the thing mapped."—Eric Temple Bell

Figure 1: Discovered underground map of New York City with 8 neighborhoods. **Left:** The photos show the top discriminative styles from each neighborhood corresponding to the color of their border. **Right:** Map produced from administrative boundaries ("Traditional", top) vs. the map produced by our method ("Underground", bottom). Traditional neighborhoods do not capture information made apparent by an underground map, e.g., discovering two far apart but similar tourist neighborhoods (red).

1. Introduction

Cities are traditionally divided into multiple neighborhoods, where neighborhood boundaries occur due to a variety of reasons, including city governance and management, geographic separation of regions (e.g., by water, hills, etc.), or historical factors (past city extension). However, a person who knows a city well often has a different notion of neighborhoods than what these boundaries provide. To take New York City as an example, even though Manhattan is divided into Downtown, Midtown, and Uptown by traditional maps, a local person may see a region in Uptown (Columbia University) as similar to a region in Downtown (NYU), since student populations live in both regions. Similarly, there are tourists south of downtown and throughout most of midtown. Lower Manhattan is not a monolithic neighborhood, but rather a culturally diverse region, e.g., encompassing both Chinatown and corporate offices near Wall Street. Often such notions of neighborhoods can be more useful than traditional neighborhood divisions, as they reveal how a neighborhood actually is perceived and experienced. We call these kind of neighborhood maps "underground" maps to differentiate them from traditional neighborhood maps.

How can we get such underground maps? While no prior work infers underground maps, past computer vision work explores "urban perception" using street-view images of buildings [29, 17, 11] and cars [12] to characterize a location. While fascinating, such glimpses of a city remain a step removed from the people who traverse it, and they are static (e.g., buildings may persist unchanged for decades, while the culture of a neighborhood evolves more rapidly). Meanwhile, directly crowdsourcing for an underground map is challenging to scale and requires city-specific expertise.

We propose to discover underground neighborhood maps from *fashion senses* observed in public social media photos. See Figure 1. The key insight is that people's clothing is a strong indicator of their personal style, interests, and current

activity, which in turn reveals a bottom-up grouping of the regions within a city. For example, people in the vicinity of a beach are likely to be found wearing beachwear, whereas people preparing to jog may wear short-sleeved shirts and shorts in warm weather. Similarly, students near a university often wear shirts with their university colors, while sports fans don the colors of their team, and others wear clothes reflecting a social cause or pop culture element that may be active in a part of a city. Unlike architecture or cars, fashion images provide dynamic information. For example, a person drives the same car in the whole city irrespective of what they plan to do in a particular neighborhood, whereas their clothes can change based on the activity (e.g., gym vs. beach vs. work). Based on these observations, we believe that fashion is an interesting yet unexplored indicator of the underground map notion.

Our approach uses the distribution of fashion styles (we call this *fashion sense*) at a place to discover an underground map. We first detect clothing attributes in 7.7M public geo-located social media photos spanning 37 cities worldwide, and then discover typical combinations of those attributes, or styles. Then, we use unsupervised clustering to detect pockets of a city that are both spatially and stylistically coherent. Finally, in addition to returning the generated neighborhood map, we devise computational measures to identify a city's most unique neighborhoods and mine for "analogical" neighborhoods between otherwise different cities (e.g., what is the "Uptown" of Bogota?). In contrast to previous work on urban perception, which requires supervision in the form of image labels for the latent property of interest [29, 17, 11, 26] our method uses no underground neighborhood labels. Figure 1 shows an example underground map created by our method for New York City, compared to a traditional administrative city map (cf. Sec 4).

There are many potential applications of underground maps. A person unfamiliar with a city could find out what neighborhoods might be suitable for them to visit, e.g., to satisfy interests in outdoor activities vs. shopping vs. tourist areas. A visitor could grasp at a glance how people typically dress in a region, e.g., when choosing attire for a restaurant. Anthropologists could leverage the mined maps to infer trends within a city and across time. A more obscure part of a city could gain positive exposure for its distinct culture. In any such case, an underground map addresses queries in ways that go beyond traditional maps.

To evaluate our approach, we also introduce two *new* non-visual benchmarks that capture the notion of underground maps. One benchmark captures how people perceive a neighborhood, while the other captures the business distribution (indirectly the activity distribution) in a neighborhood. Experiments show that our model produces accurate and coherent neighborhood regions. Further, our qualitative results and evaluation with human judges reinforce these findings and illustrate the value of fashion images as a new tool to interpret subtleties in the life of a city. Our work is the first to discover underground maps of a city and to analyze fashion *within* individual cities at a large scale. To summarize, our contributions are:

- a simple-yet-effective method that uses fashion images of a city to discover underground neighborhoods,
- two non-visual benchmarks that capture the underground neighborhood notion of 37 worldwide cities,
- methods to discover meaningful insights (e.g., uniqueness, analogies, historical expansion) from the produced underground maps.

2. Related work

Visual understanding of clothing. Computer vision techniques are actively used for fashion. Research has focused on the classification of clothing attributes [6, 5, 4, 40, 21, 25], segmenting clothing in images [37, 36, 38, 16], and product identification [8, 34, 13, 21]. There is also prior work on classifying outfits into styles, e.g., "hipster", "goth" [18]. Clothing recommendation systems address occasion-based dressing [20], location [23], or compatibility and style coverage [15]. Our work uses attribute prediction to create an embedding space for understanding the fashion sense of a city. However, the goal is not to classify fashion attributes on images, but to use the embedding to discover underground neighborhood maps for a city.

Visual style discovery. Some prior research uses visual analysis to discover styles and trends. Early work used low-level image features or mined visually distinctive patches [10, 32, 9] to discover unique architectural properties of a city. Discovering fashion styles without style supervision has also been explored [25, 14, 2]. These methods leverage attribute predictions or embeddings to discover distinct styles. Learning to detect urban tribes by using crowdsourced data has also been explored [19]. While these methods focus on discovering styles, we leverage styles to discover neighborhoods with different fashion senses.

Trend forecasting. Recent work in fashion forecasting trains temporal models to predict how a particular fashion style will rise or fall in popularity in the future. This was first addressed in [2], who forecast trends with coarse yearly predictions for a year in advance. Recent work looks at finer-grained forecasting at a weekly granularity [24, 22, 1]. These models enable discovery of unique events where people wear a specific type of clothing with an anomalously high occurrence [24] and detection of fashion influence of one city on another [1]. Unlike existing methods, our work aims to discover latent maps of cities using the fashion sense of regions *within* a city. This is a challenging task as we do not know the boundaries of neighborhoods a priori; our goal is to discover them without any supervision. We are the first to discover latent neighborhoods using fashion sense.

Urban perception. Prior work has focused on predicting geo-spatial properties within an urban environment, using the visual features at a geo-location. This includes properties like the perceived safety of cities [3, 28, 29, 11, 41], ecological properties such as snow or cloud cover [39, 35, 26], or the distance to Starbucks [17]. Advances in visual recognition have enabled sophisticated analyses, such as modeling demographics by recognizing the model of cars in StreetView [12], recognizing structures in satellite images [31], or modeling visual changes predictive of neighborhood improvements [27]. Note that all these previous methods require some form of labeling for the latent variable they are trying to perceive. For example, human annotators manually rank images based on safety [29], or demographic data of poverty is needed to predict poverty in an unseen region [31]. In contrast, our approach does not use any such labels for the segmentations it produces. To our knowledge, we are the first to perceive such a factor without direct supervision, and the first to address the problem of underground maps.

3. Method

Our goal is to segment the map of a city into regions based on the fashion sense in the region. First we provide background on the problem, and we then discuss our method. Fig. 2 overviews our approach.

3.1. Dataset and Style Discovery

To understand fashion sense within a neighborhood of a city, we aim to understand the clothing in images of people in that neighborhood. Therefore, we need a dataset that can capture how people dress at different locations in a city. To get a real-world glimpse (see "dataset and possible biases" for some limitations) of what people are wearing across a city, we use images sampled from social media platforms; specifically, we use the 7.7M images from the GeoStyle dataset [24] from Instagram and Flickr.

For a city, let $\{I_i\}$ be the set of images of people. We also know the geolocation tagged to each of these images, $\boldsymbol{l_i} \in \mathbb{R}^2$ for an image I_i. Following the method in [24], we first train a representation by learning to classify basic clothing attributes. The attributes consist of a variety of fashion properties, like clothing type (suit, t-shirt, dress etc.), presence of accessories (sunglasses, necktie), color of clothing (red, blue), among others. We use these attribute annotations on a small dataset of 27k images to train a multi-task GoogLeNet[33] CNN, where each head classifies a particular attribute. Using the multi-task CNN, we can predict the attributes for new images. We denote an attribute vector for an image I_i by $\boldsymbol{a}_i \in \mathbb{R}^A$, where A is the number of attributes. Note that these attributes capture the properties of clothes, not the identity of people wearing them.

We then learn a set of global styles that capture combinations of basic attributes by leveraging all the images in the dataset. We cluster the feature representation (from the penultimate layer of the CNN) of all the images, using a Gaussian Mixture Model (GMM) with $K = 400$ components. We denote the style prediction by this model for an image I_i by $s_i \in [1, K]$. The number of styles is set to yield visually coherent styles, following [2, 14, 24, 1]. Note that while we use clothing attributes to learn a good style representation, unlike [3, 28, 29, 11, 41] our model does not require any labeling for the latent variable of interest (here, neighborhood membership) and is thus **unsupervised**.

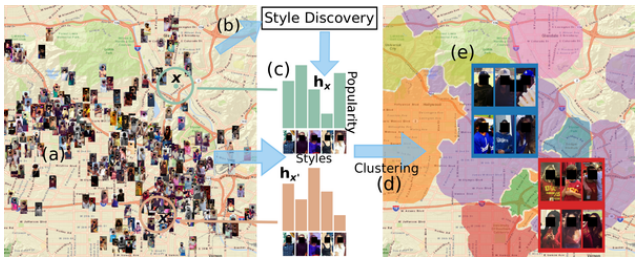

Figure 2: Pipeline to discover fashion neighborhoods. (a) Input: large set of geo-tagged fashion images. (b) We perform unsupervised style discovery to get styles. (c) Using the styles and geo-tagged images, we describe a location x using its style histogram h_x. (d) Finally, we cluster locations using the descriptors. (e) Neighborhoods can be visualized by looking at their top styles.

3.2. Featurizing a Geolocation

We want to characterize locations in a city by the fashion sense of their surroundings. The fashion in an image I_i with location l_i can describe a location l_i. However, a location is not described by a single style but a distribution over styles. Therefore, we describe the fashion sense of a particular location using the images in its vicinity. Specifically, to describe a geolocation $\boldsymbol{x}$ we select all images within a radius r from that geolocation. This formulation lets us capture the distribution of fashion in a surrounding area. Additionally, it ensures the fashion distribution changes smoothly from one location to another, and implicitly enforces that nearby locations should belong to the same neighborhood.

Let $T(\boldsymbol{x})$ be the set of images that describes location $\boldsymbol{x}$:

$$T(\boldsymbol{x}) = \{I_i : ||\boldsymbol{l}_i - \boldsymbol{x}||_2 < r\}. \quad (1)$$

The images' distribution over the different styles can describe the location. For example, a region near a beach would have a higher distribution over the styles that are unique to a beach (board shorts, tank tops, etc.). Therefore, we compute a histogram $\boldsymbol{h_x} \in \mathbb{R}^K$ over the K styles of the images in $T(\boldsymbol{x})$ to describe location $\boldsymbol{x}$. Since the sampled images will be biased towards a direction we compute an unbiased location for each $\boldsymbol{x}$ (see Supplementary).

There are tradeoffs in choosing the radius r. A small radius would let us create good local features, but would

result in a low confidence histogram as there would be very few sampled images. A large radius would result in a high confidence histogram, but would capture a larger portion of the map. We set the radius so that the ratio of intersection over union of adjacent sampling regions is close to 0.5.

3.3. Sampling Locations and Clustering

Next, we use this local featurization at every point to segment the map into regions. We cluster regions based on similarities in their histogram descriptors.

Locations $\boldsymbol{x_{ij}}$ are sampled uniformly at distance d from a 2D grid over the map of a city. We sample images $T(\boldsymbol{x_{ij}})$ around $\boldsymbol{x_{ij}}$ for all locations, and obtain the histogram descriptor $\boldsymbol{h}_{\boldsymbol{x}_{ij}}$. We use K-means to cluster the histogram descriptors to get a label for each location. Since the L1 norm for histograms is 1, we use L1 instead of the standard Euclidean distance when performing the M-step of K-means.

This grouping step is simple but effective—more effective than other more elaborate variants we explored (discussed in Supp.). In the next three sections we discuss how these discovered neighborhoods can be used for analysis and applications like finding unique neighborhoods of a city, or finding similar and "analogical" neighborhoods across cities.

3.4. Finding Unique Neighborhoods

Having computed the neighborhoods, we can calculate which neighborhoods have the most unique fashion sense. A unique neighborhood is defined in our framework as a neighborhood most distinct from all other neighborhoods in that city. Each discovered neighborhood is described by a histogram descriptor $\boldsymbol{h_{n,c}}$ that is created by aggregating images in that neighborhood. We use distances between these descriptors to identify the most unique neighborhood, namely, a unique neighborhood has the maximum L1 distance from its most similar neighborhood in the same city:

$$n_{\text{unique}} = \arg\max_{n} \min_{m \in N, m \neq n} ||\boldsymbol{h_{n,c}} - \boldsymbol{h_{m,c}}||_1. \quad (2)$$

We also sort neighborhoods by this distance over all the cities, so that we can rank the most unique neighborhoods with their cities.

3.5. Finding Similar Neighborhoods Across Cities

Inspired by applications noted in Sec. 1, we use the discovered neighborhoods to detect similar neighborhoods from two different cities. To know which neighborhood of a city is like some other neighborhood in another city, we find the L1 distances between histogram descriptors of all the neighborhoods of all the cities and sort them by this distance.

3.6. Finding Neighborhood Analogies

Finally, we introduce a method to identify *neighborhood analogies*. The above method is successful in finding similar neighborhoods within cities having similar weather and culture. However, if the weather and culture are significantly different for two cities, histogram distances between any of their neighborhoods will be large and hence less meaningful. To determine neighborhood analogies, instead of directly measuring distances between neighborhoods, we encode each neighborhood in the context of its city. By measuring similarities of neighborhoods using this contextual encoding, we recover analogical pairs of neighborhoods across cities. An example of such a pair could be *Bondi beach* : Sydney :: *Coney Island* : NYC, both with popular beaches. Since we are encoding *Bondi beach* with respect to Sydney and *Coney Island* with respect to NYC, the similarity between these two would be a measure of this analogy.

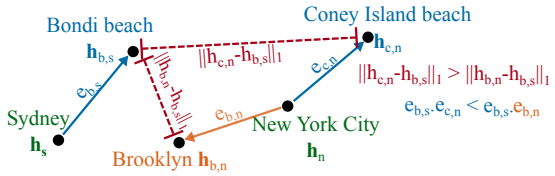

Figure 3: Illustration of the usefulness of our analogy-inspired encoding. The two beaches have similar relationships to their city, hence they should be more similar than other neighborhoods—even though their absolute similarity may be low.

Each neighborhood has a histogram descriptor $\boldsymbol{h}_{n,c}$, and the city has an aggregate histogram descriptor $\boldsymbol{h}_c$. We define the contextual encoding of a neighborhood n with respect to its city c as:

$$e_{n,c} = \text{sgn}(\boldsymbol{h}_{n,c} - \boldsymbol{h}_c). \quad (3)$$

This encoding contains information about which styles are more popular with respect to other neighborhoods in the city. It ignores the magnitudes of relative style popularity and only considers direction, providing invariance to exact style popularities. We measure the cosine distance between pairs of neighborhoods across cities to find the analogically similar pairs. Figure 3 illustrates how contextual encoding can better capture analogies. If cities are geographically far apart, there might be a shift in the overall distribution, and a contextual encoding (blue) would produce better results than a non-contextual encoding (red). Note that if the two cities are similar, both will measure similar quantities.

Dataset and possible biases. We implement our method on the 7.7M-image GeoStyle dataset [24] where each city has 175k images. For experiments, we use $r = 0.02°$ and $d = 0.01°$. Exploration of the impact of these hyperparameters is done in the Supplementary. We employ GeoStyle because it is the largest publicly available dataset of its kind and offers a real-world glimpse of what people wear across the world. While powerful, it has certain limitations. Images in Instagram (or from any other social media platform) will have sampling biases. For example, Instagram is known to be most popular amongst young users. Given that the

dataset is biased to particular age groups, the neighborhoods we discover are also going to be influenced by these age groups. The data also focuses on cities rather than rural areas. Also, a region with tourist attractions is likely to have a higher fraction of photos taken by tourists as compared to non-tourist people. Hence, the fraction of tourist vs. non-tourist images will likely not reflect the true density of tourist vs. non-tourist populations, and we can expect our method to be influenced by tourists and more photogenic places.

4. Results

We quantitatively evaluate our method's ability to discover underground maps. Additionally, we look at qualitative results and evaluate the methods with human judgements. See Supplementary and video for many more examples.

4.1. Benchmarks

To judge our method's ability to segment the city into neighborhoods, we need to evaluate it against some notion of ground truth. Note that it is impossible to obtain absolute ground truth information of underground maps, as there is no single concrete definition of what should be the property constituting underground maps. Instead we consider multiple approximations of underground maps for evaluation.

We create two such benchmarks. The first is obtained using an external, publicly crowdsourced platform called HoodMaps[1] and looks at how people from a city perceive different neighborhoods. This benchmark captures subjective impressions. The second is created using business densities of different types using OpenStreetMap.[2] This benchmark captures objective measures. Both segment cities into regions based on differences in their properties.

HoodMap (HM) Benchmark: We create the HoodMap (HM) Benchmark using information from the crowdsourced public platform called HoodMaps. The users on this platform can label regions of a city. The service aggregates the votes to provide the majority-voted label for a region. Figure 4 (Left) shows the map of Chicago with the 6 associated labels. HoodMaps has a granularity of 0.01° along both latitude and longitude. We collect this information for all 37 of the 44 GeoStyle cities with enough data (see Supp), comprising votes from more than 30k people.

The HM annotations are based on perceptions of people, and hence they can capture a notion of neighborhoods beyond geographic boundaries. For the 37 cities, a region is voted on by more than 70 voters on average, and more than 55% of voters agree on a single label out of the 6 (chance would be 16%). This relatively high number of votes and agreement indicates the labelling is consistent. However, HoodMaps does come with certain limitations. The coarse

[1] https://hoodmaps.com/

[2] https://www.openstreetmap.org/

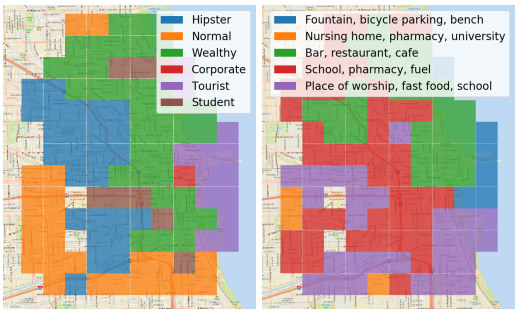

Figure 4: We analyze our discovered neighborhoods against two complementary non-visual sources for underground maps. **Left:** HM benchmark labels and neighborhoods dividing Chicago into 6 neighborhoods based on how people perceive them. **Right:** BD benchmark neighborhoods dividing the city into 5 regions based on its business/amenity distributions. The legend shows 3 businesses that are more frequent in that particular neighborhood.

label set was selected by HoodMaps (not us) and is not necessarily complete for all categories of interest. The labels may also display certain stereotypes that conflict with the ideal underground map. For example, one of the labels is "wealthy", yet our goal is to divide cities on popular activities or interests, *not* socio-economic status. There is also a possibility of sampling bias amongst people who choose to visit HoodMaps.com and vote for these labels. In short, while the data is a useful non-visual source for perceived neighborhoods, we also can expect our image-based results to deviate meaningfully from those boundaries, possibly in ways that challenge common stereotypes.

Business Distribution (BD) Benchmark: While the HoodMap benchmark captures how people perceive a neighborhood of a city, it need not capture what activities are present in a neighborhood. Therefore, we create the Business Distribution (BD) Benchmark that captures the distribution of business types across a city. We use OpenStreetMap to get geolocated businesses, for total of 1,446 different business types and 1.6M businesses/amenities. The distribution of the frequency of business types is long-tailed, so we consider businesses types with frequency at least 50 (we find 154 such types). The distribution of businesses gives a more objective measure of a neighborhood in contrast to HoodMaps. For example, a region with a higher distribution of pubs/nightclubs is likely catering to a different crowd than a region with libraries/schools. Similarly, a large density of museums indicates a region popular amongst tourists. To create regions over maps using business density, we follow our method from Sec. 3. More details are in Supplementary. Figure 4 (Right) shows the BD map of Chicago, along with the amenities/businesses that are more frequent in each region.

In short, the HM and BD benchmarks capture complementary notions of an underground map. The former captures how people perceive a neighborhood, whereas the latter captures the activities one can do in a neighborhood. We will

publicly release these benchmarks to facilitate future work.

4.2. Baselines

We evaluate with the following set of baselines.

Random: is a naïve baseline, where every point of interest is assigned a random label.

Proximity: clusters on geographical proximity instead of style histograms. This baseline uses the uniformly sampled locations (lat. and long. pair) as the feature for clustering.

Proximity+Image Density (PID): clusters on x_{ij} and hence leverages additional information about the image density at different locations. This baseline is stronger than using proximity alone, as image density can tell a lot about a neighborhood, e.g., a residential area is likely to have lower density, unlike a tourist area with lots of photos.

Caption: clusters image captions. Image captions are clustered using aggregated GloVe vectors [30] for each city and we use histograms over these clusters as features. This baseline is similar to our method but uses a non-visual modality instead of visual or fashion-specific cues.

Full image: sees if the image background provides more useful information about neighborhoods than just looking at the clothing features. It uses features from a ImageNet pre-trained CNN instead of fashion features to create style histograms. It captures global information of people and their background, such as street-view buildings, vegetation, etc. For fair comparison we use GoogLeNet, same as for our fashion attributes (cf. Sec. 3.1).

Administrative boundaries (Admin): represents traditional maps based on government issued boundaries. It uses ordinance maps for the 8 GeoStyle cities for which we could find publicly available data by the city.[3] These boundaries are fine-grained, so we greedily merge them based on proximity to match the granularity of the other baselines.

4.3. Example Maps and Quantitative Evaluation

Figs. 1 and 5 show example maps for NYC, Seattle, Delhi, and Bangkok, along with the Admin baseline's map when available. These exemplify some intuitive and some unexpected findings by our model. For example, it discovers traditional vs. Westernized neighborhoods: in Delhi, people in the southern neighborhood are seen in Western clothing more than in the north, explainable by the fact that the city expanded from north to south as it grew; in Bangkok, we see traditional vs. new neighborhoods (pink, yellow). Such revealing visualizations offer a new tool to tourists, visitors, or even anthropologists to make such discoveries.

We first evaluate how much neighborhoods discovered by our method align with the neighborhoods of the benchmarks. We use 3 unsupervised clustering metrics: (i) Normalized Mutual Information (NMI) captures the mutual information

[3]See Supplementary for list of cities. Example for Los Angeles: [7]

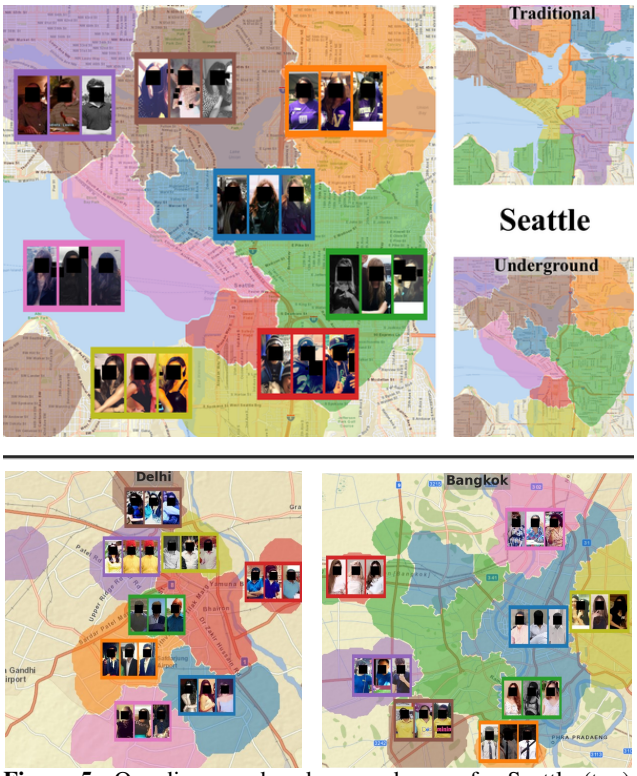

Figure 5: Our discovered underground maps for Seattle (top), Delhi (bottom left), and Bangkok (bottom right). Please see Supp. for many more examples and discussion.

between the produced maps and benchmarks. (ii) Purity captures the maximum precision of the produced maps w.r.t. the benchmarks. (iii) Mean Maximum Intersection over Union (MMIoU) measures the mean IoU of each patch's best matching patch. We report performance where the number of clusters equals the number of labels in the benchmarks (see Supplementary for sweeps over cluster numbers). For HoodMaps, this number is 6, and for Business Distribution this number is different for different cities, based on the number of clusters produced by affinity propagation.

Table 1 shows the results. The left side compares our method to all baselines except Admin aggregated over all 37 cities; the right side compares all methods on the 8 cities for which the baseline Admin is applicable. Our method outperforms all baselines on both benchmarks on all the metrics for the "all cities" test, and also outperforms the Admin baseline for every city it is available except for one metric on BD.

All cities. Table 1 (left) reveals a few things. First, the *proximity+image density* baseline performs well and is a very strong baseline. This shows that the similarity of a region is strongly affected by the image density. Second, our performance against the *caption* baseline shows that visual fashion sense provides more useful cues than the manually written captions of the images. While *caption* does

	All Cities						Cities with Administrative Boundaries Available					
	HM Benchmark			BD Benchmark			HM Benchmark			BD Benchmark		
Method	NMI	Purity	MMIoU	NMI	Purity	MMIoU	NMI	Purity	MMIoU	NMI	Purity	MMIoU
Random	0.079	0.464	0.172	0.092	0.359	0.173	0.084	0.431	0.171	0.090	0.545	0.170
Proximity	0.225	0.542	0.249	0.342	0.559	0.325	0.249	0.546	0.269	0.288	0.609	0.281
PID	0.242	0.550	0.262	0.353	0.579	0.336	0.277	0.597	**0.281**	0.303	0.665	**0.294**
Admin	-	-	-	-	-	-	0.235	0.570	0.256	0.282	0.686	0.260
Caption	0.222	0.607	0.235	0.332	0.561	0.313	0.207	0.644	0.215	0.299	0.731	0.278
Full image	0.248	0.573	0.244	0.331	0.589	0.284	0.271	0.631	0.270	0.312	0.735	0.262
Ours	**0.260**	**0.635**	**0.272**	**0.369**	**0.597**	**0.339**	**0.291**	**0.652**	**0.281**	**0.323**	**0.742**	0.281

Table 1: Comparison of our method against baselines on both the HM and BD benchmarks. The first six columns show results on all the cities. The last six columns show the results on cities where administrative boundary (Admin) data is available. Our method performs better than all the baselines and all the metrics except for one. Our gains over the Admin baseline accentuate how traditional maps (e.g., by city ordinances) are distinct from the underground perceived maps of a city.

	Wea.	Hip.	Tou.	Stu.	Nor.	Cor.
PID	0.297	0.281	0.247	0.223	0.256	0.171
Ours	0.293	0.289	0.265	0.243	0.258	0.166

Table 2: Per class accuracy on HoodMaps (MMIoU). Note that the labels were chosen by HoodMaps, and may be associated with stereotypes that we do not aim to discover (see Sec. 4.1).

contain useful information, it also contains information not necessarily related to location. Third, the *full image* baseline performs worse than our method. This shows that fashion specifically is indeed a useful factor for understanding and discovering neighborhoods, beyond the surrounding buildings, weather, etc. as captured in the full image and often used in existing urban perception tasks.

Cities with administrative boundaries. Table 1 (right) rescores all baselines and our method for the 8 cities where Admin is possible. We see Admin is no better than the proximity baseline. This accentuates that the "traditional" measure of a neighborhood is different from the "underground" notion. Figure 1 (bottom right) shows the Admin boundaries for New York City for 8 regions: Admin's manually demarcated neighborhoods cannot find distant similar regions. For example, whereas we discover regions popular among tourists (red, top right) that are similar to each other, a "traditional" neighborhood map of a city will not find such similarities. Underground maps offer significantly different information than traditional maps.

Per-class performance. Next, to understand which underground neighborhood types are best revealed by fashion, we analyze the per-class performance. Table 2 shows the results against the best baseline, PID.[4] Our method finds it easier to discover tourist and student classes, suggesting it is possible to determine these neighborhoods by looking at changes of particular styles. Interestingly, the image density itself (PID) reveals cues about hipster and wealthy areas. Our method finds it difficult to discover corporate neighborhoods. We believe this is because fewer people are posting images of themselves from a corporate environment.

[4]Here we measure MMIoU, as purity and NMI are aggregate measures.

Figure 6: Two unique neighborhoods. Los Angeles (Left), Bogota (Right). Each row shows the distinctive style of the neighborhood.

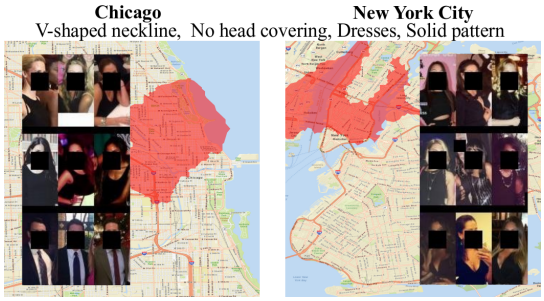

Figure 7: Most similar neighborhood in Chicago and NYC. Text on top shows their most popular inferred attributes.

4.4. Qualitative Results

Having showed that the discovered neighborhoods successfully capture underground notions of a city, we now show how we can use these discovered neighborhoods to get useful information about a city and relationships between cities (cf. Sec. 3.4 to 3.6). Due to space limits, we display some examples here, then provide many more examples in the Supplementary, together with subtle or unexpected insights suggested by our method's discoveries.

Most unique neighborhoods. Figure 6 shows some of the unique neighborhoods discovered by our method (Sec. 3.4). First, we find a neighborhood around Dodgers stadium in Los Angeles where a very high number of people can be seen wearing blue colored tops and baseball hats (the team color). The second neighborhood in Bogota is popular among tourists for hiking and outdoor activities, where

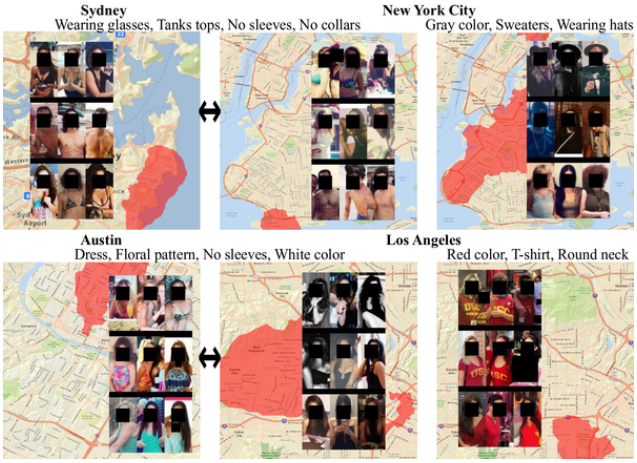

Figure 8: Analogy neighborhoods between Sydney and NYC (top) and Austin and Los Angeles (bottom). In each row, the first two images show the analogy obtained with our contextual encoding, and the rightmost image shows the (baseline) similar neighborhood computed without the contextual encoding. Text shows most popular attributes. Our model surfaces answers to nuanced questions like "How do people dress in the bar scenes of Austin vs. LA?"

people can be seen wearing glasses and winter clothing.

Similar neighborhoods across cities. Figure 7 shows similar neighborhoods for Chicago and NYC found using our method in Sec. 3.5. They show a relatively high fraction of people in party wear. Both the neighborhoods are popular places for nightclubs as confirmed by the BD data.

Neighborhood analogy. Figure 8 shows example analogical pairs of neighborhoods found across regions using the proposed contextual encoding (Sec. 3.6). Our method is able to discover similar neighborhoods across continents with significantly different weather and culture. With contextual encoding, we find two neighborhoods with popular beaches in Sydney and New York City (NYC). Because the histogram shift is too large from Sydney to NYC, if we do not use the proposed contextual encoding an incorrect neighborhood is found (see Figure 8 (top right)). The difference is further exemplified by the discriminative cluster images shown from the two regions, which highlight beachwear. Bottom row shows analogical neighborhoods between Austin and Los Angeles. Dresses can be seen in the analogical pair as opposed to the neighborhood found without context (right).

4.5. Experiments with Human Judges

A denizen of a city has a deeper understanding of its neighborhoods. Does our technique match what that person would say? Using Mechanical Turk to reach "locals" (see Supp.), we display a set of images from discovered neighborhoods, and ask the worker to select the point on the map they think the images come from. Images from neighborhoods of both our method and the strongest baseline (PID) are shown. We measure both raw accuracy and the Area Normalized Accuracy (ANA), which is weighted by the inverse of the area of the region. The latter accounts for the fact that the areas of neighborhoods produced by both methods could be unequal, and clicking in the correct region by chance is more probable for the larger region.

Method	Acc.	ANA	Confidence
Proximity+Image Density	14.28	15.00	0.57
Ours	**24.90**	**27.50**	**0.61**

Table 3: Human subjects are more frequently able to identify where the neighborhood style comes from when using our neighborhoods.

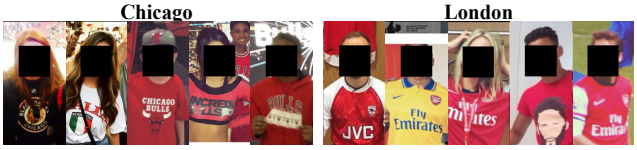

Figure 9: Random samples from two sets of neighborhoods for which all workers confidently point to the correct location on maps.

Table 3 shows the results, accumulated over all cities and judged by at least three workers. Our method does significantly better than the best PID baseline. The confidence values (0: no confidence, 1: very confident) show users were more confident in figuring out the regions when images come from our method. This result is exciting as it suggests that the neighborhoods discovered automatically by our method are recognizable as coherent by people who know the cities. This is evidence our method produces indicative properties about neighborhoods that can be verified by the people from the city. Figure 9 shows examples for which all workers confidently pointed to the correct place. Both neighborhoods contain famous stadiums of the cities' teams (Blackhawks and Arsenal) and are easily identified.

5. Conclusion

We introduced underground fashion maps for cities and proposed the first method to create them. This is an original problem for vision+fashion. We show how insights mined from large-scale clothing images can solve it, with rigorous analysis against two new benchmarks for how people perceive a city. The benchmarks will be shared publicly to allow continued work in this area. Our work fits in with the long tradition of vision research to publish exploratory uses of big visual data, where known vision techniques are used to achieve something completely new [17, 35, 29, 41, 26, 11, 2, 24, 1].

We demonstrate various novel uses for the maps, including finding distinctive neighborhoods and analogies across cities, with the latter finding associations between cities that are missed by the more straightforward approach. Our maps capture the sense of a neighborhood better than current administrative maps or other image baselines, and our user study reveals the value of fashion images as a new tool to interpret the life of a city. In future work, we plan to explore how underground maps evolve over time as styles change.

Acknowledgements. This work was done as part of an internship at Facebook AI Research. We also thank TCS.

References

[1] Ziad Al-Halah and Kristen Grauman. From Paris to Berlin: Discovering fashion style influences around the world. In *CVPR*, 2020.

[2] Ziad Al-Halah, Rainer Stiefelhagen, and Kristen Grauman. Fashion forward: Forecasting visual style in fashion. In *ICCV*, 2017.

[3] S.M. Arietta, A.A. Efros, R. Ramamoorthi, and M. Agrawala. City Forensics: Using visual elements to predict non-visual city attributes. *IEEE Trans. Visualization and Computer Graphics*, 20(12), Dec 2014.

[4] Lukas Bossard, Matthias Dantone, Christian Leistner, Christian Wengert, Till Quack, and Luc Van Gool. Apparel classification with style. In *Proc. Asian Conf. on Computer Vision*, 2013.

[5] Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Describing people: Poselet-based attribute classification. In *ICCV*, 2011.

[6] Huizhong Chen, Andrew Gallagher, and Bernd Girod. Describing clothing by semantic attributes. In *ECCV*, 2012.

[7] City of Los Angeles Open Data. LA Times neighborhood boundaries.

[8] Wei Di, C. Wah, A. Bhardwaj, R. Piramuthu, and N. Sundaresan. Style finder: Fine-grained clothing style detection and retrieval. In *Proc. CVPR Workshops*, 2013.

[9] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Mid-level visual element discovery as discriminative mode seeking. In *NeurIPS*, 2013.

[10] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei A. Efros. What makes Paris look like Paris? *SIGGRAPH*, 31(4), 2012.

[11] Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, and César A Hidalgo. Deep learning the city: Quantifying urban perception at a global scale. In *ECCV*, 2016.

[12] Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. *Proc. National Academy of Sciences*, 2017.

[13] M Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C Berg, and Tamara L Berg. Where to buy it: Matching street clothing photos in online shops. In *ICCV*, 2015.

[14] Wei-Lin Hsiao and Kristen Grauman. Learning the latent "look": Unsupervised discovery of a style-coherent embedding from fashion images. In *ICCV*, 2017.

[15] Wei-Lin Hsiao and Kristen Grauman. Creating capsule wardrobes from fashion images. In *CVPR*, 2018.

[16] Menglin Jia, Mengyun Shi, Mikhail Sirotenko, Yin Cui, Claire Cardie, Bharath Hariharan, Hartwig Adam, and Serge Belongie. Fashionpedia: Ontology, segmentation, and an attribute localization dataset. *CoRR*, 2020.

[17] Aditya Khosla, Byoungkwon An, Joseph Lim, and Antonio Torralba. Looking beyond the visible scene. In *CVPR*, 2014.

[18] M. Hadi Kiapour, Kota Yamaguchi, Alexander C. Berg, and Tamara L. Berg. Hipster wars: Discovering elements of fashion styles. In *ECCV*, 2014.

[19] Iljung S Kwak, Ana Cristina Murillo, Peter N Belhumeur, David J Kriegman, and Serge J Belongie. From bikers to surfers: Visual recognition of urban tribes. In *BMVC*. Citeseer, 2013.

[20] Si Liu, Jiashi Feng, Zheng Song, Tianzhu Zhang, Hanqing Lu, Changsheng Xu, and Shuicheng Yan. Hi, magic closet, tell me what to wear! In *Proc. Int. Conf. on Multimedia*, 2012.

[21] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016.

[22] Yunshan Ma, Yujuan Ding, Xun Yang, Lizi Liao, Wai Keung Wong, and Tat-Seng Chua. Knowledge enhanced neural fashion trend forecasting. In *ICMR*, 2020.

[23] Yunshan Ma, Xun Yang, Lizi Liao, Yixin Cao, and Tat-Seng Chua. Who, where, and what to wear? extracting fashion knowledge from social media. In *Proc. Int. Conf. on Multimedia*, 2019.

[24] Utkarsh Mall, Kevin Matzen, Bharath Hariharan, Noah Snavely, and Kavita Bala. GeoStyle: Discovering fashion trends and events. In *ICCV*, 2019.

[25] Kevin Matzen, Kavita Bala, and Noah Snavely. Streetstyle: Exploring world-wide clothing styles from millions of photos. *CoRR*, 2017.

[26] Calvin Murdock, Nathan Jacobs, and Robert Pless. Building dynamic cloud maps from the ground up. In *ICCV*, 2015.

[27] Nikhil Naik, Scott Duke Kominers, Ramesh Raskar, Edward L Glaeser, and César A Hidalgo. Computer vision uncovers predictors of physical urban change. *PNAS*, 2017.

[28] Nikhil Naik, Jade Philipoom, Ramesh Raskar, and César Hidalgo. Streetscore: Predicting the perceived safety of one million streetscapes. In *Proc. CVPR Workshops*, 2014.

[29] Vicente Ordonez and Tamara L. Berg. Learning high-level judgments of urban perception. In *ECCV*, 2014.

[30] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[31] Anthony Perez, Christopher Yeh, George Azzari, Marshall Burke, David Lobell, and Stefano Ermon. Poverty prediction with public landsat 7 satellite imagery and machine learning. *NeurIPS*, 2017.

[32] Saurabh Singh, Abhinav Gupta, and Alexei A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012.

[33] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[34] S. Vittayakorn, K. Yamaguchi, A.C. Berg, and T.L. Berg. Runway to Realway: Visual analysis of fashion. In *WACV*, 2015.

[35] Jingya Wang, Mohammed Korayem, and David J. Crandall. Observing the natural world with flickr. In *ICCV Workshops*, 2013.

[36] K. Yamaguchi, M.H. Kiapour, and T.L. Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *ICCV*, 2013.

[37] K. Yamaguchi, M.H. Kiapour, L.E. Ortiz, and T.L. Berg. Parsing clothing in fashion photographs. In *CVPR*, 2012.

[38] Wei Yang, Ping Luo, and Liang Lin. Clothing co-parsing by joint image segmentation and labeling. In *CVPR*, 2014.

[39] Haipeng Zhang, Mohammed Korayem, David J. Crandall, and Gretchen LeBuhn. '. In *WWW*, 2012.

[40] Ning Zhang, Manohar Paluri, Marc'Aurelio Rantazo, Trevor Darrell, and Lubomir Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *CVPR*, 2014.

[41] Bolei Zhou, Liu Liu, Aude Oliva, and Antonio Torralba. Recognizing city identity via attribute analysis of geo-tagged images. In *ECCV*, 2014.