Sustainable Cities and Society 79 (2022) 103630

Contents lists available at ScienceDirect

Sustainable Cities and Society

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

CitySurfaces: City-scale semantic segmentation of sidewalk materials

Maryam Hosseini [a,b,*], Fabio Miranda [c], Jianzhe Lin [a], Claudio T. Silva [a]

[a] *New York University, United States of America*
[b] *Rutgers University, United States of America*
[c] *University of Illinois at Chicago, United States of America*

ARTICLE INFO

Keywords:
Sustainable built environment
Surface materials
Urban heat island
Semantic segmentation
Sidewalk assessment
Urban analytics
Computer vision

ABSTRACT

While designing sustainable and resilient urban built environment is increasingly promoted around the world, significant data gaps have made research on pressing sustainability issues challenging to carry out. Pavements are known to have strong economic and environmental impacts; however, most cities lack a spatial catalog of their surfaces due to the cost-prohibitive and time-consuming nature of data collection. Recent advancements in computer vision, together with the availability of street-level images, provide new opportunities for cities to extract large-scale built environment data with lower implementation costs and higher accuracy. In this paper, we propose CitySurfaces, an active learning-based framework that leverages computer vision techniques for classifying sidewalk materials using widely available street-level images. We trained the framework on images from New York City and Boston and the evaluation results show a 90.5% mIoU score. Furthermore, we evaluated the framework using images from six different cities, demonstrating that it can be applied to regions with distinct urban fabrics, even outside the domain of the training data. CitySurfaces can provide researchers and city agencies with a low-cost, accurate, and extensible method to collect sidewalk material data which plays a critical role in addressing major sustainability issues, including climate change and surface water management.

1. Introduction

As urban areas expand around the world, more impervious surfaces replace the natural landscape, creating significant ecological, hydrological, and economic disruptions (Arnold & Gibbons, 1996; Chithra, Nair, Amarnath, & Anjana, 2015). Choosing the right material to cover city surfaces has become a critical issue in mitigating the adverse effects of increased anthropogenic activities. Historically, local availability, cost, strength, and aesthetics were the main factors influencing the choice of surface pavements (Lay, Metcalf, & Sharp, 2020; Tillson, 1900). The advent of asphalt and, later, concrete changed the face of cities. The longevity and durability coupled with relatively low production and installation costs made them the pavements of choice. However, as it was later revealed, these benefits came with huge environmental burdens (Van Dam et al., 2015).

One of the concerning environmental impacts of impervious surfaces is the sharp rise in urban temperature compared to its neighboring rural areas – a phenomenon called Urban Heat Island (UHI) effect (Oke, 1982). UHI, which poses serious challenges to public health, ecological environment, and urban liveability (Estoque, Murayama, & Myint, 2017), is shown to be directly associated with surface characteristics, such as thermal performance and reflectivity. It can influence microclimates within the city by absorbing more diurnal heat and emitting that into the atmosphere at night (Nwakaire, Onn, Yap, Yuen, & Onodagu, 2020; Takebayashi & Moriyama, 2012; Wu, Sun, Li, & Yu, 2018). Natural surfaces and vegetation increase the amount of evapotranspiration and decrease the overall temperature and create a cool island effect (Amati & Taylor, 2010; Du et al., 2017). Reflective/high-albedo materials are also known to decrease UHI (Akbari, Menon, & Rosenfeld, 2009; Santamouris, 2013; Santamouris, Synnefa, & Karlessi, 2011; Zhu & Mai, 2019). Hence, the spatial distribution of land cover has a strong impact on the surface temperature (Chen & Zhang, 2017). Surface material also impact the water runoff and increase the risk of flooding. Sidewalks and roads form the main part of the urban ground surfaces. Today, the majority of the sidewalks are covered with impermeable materials which prohibit the infiltration of the water into the underlying soil, increase both the magnitude and frequency of surface runoffs (Bell, Tague, & McMillan, 2019; Shuster, Bonta, Thurston, Warnemuende, & Smith, 2005), reduce the groundwater recharge, and negatively impact the water quality. The excessive use of impervious surfaces is shown to be the primary cause of the Combined Sewer Overflows (CSOs), which can lead to massive pollution of natural bodies of water and street flooding (Joshi,

[*] Correspondence to: The Visualization, Imaging and Data Analysis Research Center (VIDA), 370 Jay St, Brooklyn, NY 11201, United States of America.
E-mail addresses: maryam.hosseini@nyu.edu (M. Hosseini), fabiom@uic.edu (F. Miranda), jl12396@nyu.edu (J. Lin), csilva@nyu.edu (C.T. Silva).

https://doi.org/10.1016/j.scs.2021.103630
Received 16 June 2021; Received in revised form 17 December 2021; Accepted 17 December 2021
Available online 6 January 2022

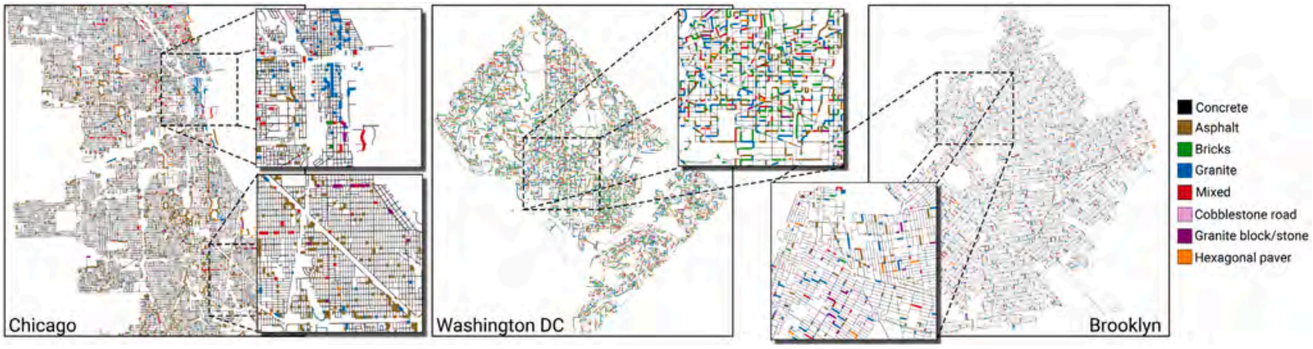

Fig. 1. Using CitySurfaces to map the dominant surface material in Chicago, Washington DC, and Brooklyn (not part of our training data). Segments where the dominant material differs from concrete are drawn using a thicker line.

Leitão, Maurer, & Bach, 2021). Aside from the mentioned impacts, sidewalk pavements can also lead to public health hazards such as outdoor falls, or pose a barrier to walkability and accessibility of public spaces, specifically for the more vulnerable population and wheelchair users (Aghaabbasi, Moeinaddini, Shah, Asadi-Shekari, & Kermani, 2018; Clifton, Smith, & Rodriguez, 2007; Talbot, Musiol, Witham, & Metter, 2005). Studies show that uneven surfaces, indistinguishable surface colors, and low-friction materials contribute to the high incidence of outdoor falls in elderly populations (Chippendale & Boltz, 2015; Thomas, Gardiner, Crompton, & Lawson, 2020).

Despite the substantial economic, environmental, public health, and safety implications of sidewalk pavements (Estoque et al., 2017; Muench, Anderson, & Bevan, 2010; Van Dam et al., 2015), most cities, even in industrialized economies, still lack information about the location, condition, and paving materials of their sidewalks (Deitz, Lobben, & Alferez, 2021). The lack of data creates barriers to understanding the real extent of the environmental and social impacts of using different materials and inhibits our ability to take a complex system approach to sustainability assessment (Van Dam et al., 2015). For instance, studies show a significant intra-urban variability of the urban thermal environment due to the street-level heterogeneity of paving materials (Agathangelidis, Cartalis, & Santamouris, 2020). However, the data scarcity makes it challenging to measure this variability across different neighborhoods and consequently, impedes the development of a sustainable and resilient mitigation response plan (Akbari & Rose, 2008; Li, Zhou, & Ouyang, 2013; Yang et al., 2019). In the absence of fine-scale data, studies mainly rely on remote sensing images; however, the high-resolution aerial images are both spatially and temporally sparse (Zhang, Odeh, & Han, 2009), requiring researchers to use a variety of data aggregation and extrapolation techniques to fill in the missing data, which can lead to high bias and hurt the validity of the final results.

Collecting comprehensive and fine-scale sidewalk data using conventional methods is time-consuming and cost-prohibitive. Recent technological innovations in data collection opened new frontiers for research on public space and pedestrian facilities, creating opportunities to track features of interest at higher temporal frequencies and more granular geographic scales (Glaeser, Kominers, Luca, & Naik, 2018). The use of street-level images in urban analysis has gained popularity since the introduction of Google Street View (GSV) (Anguelov et al., 2010) and Microsoft Street Slide (Kopf, Chen, Szeliski, & Cohen, 2010), services that provide panoramic images captured by cameras mounted on a fleet of cars. Concurrently, developments in machine learning and computer vision applied to these new datasets have enabled novel research directions to measure the "unmeasurable" in urban built environments (Ewing & Handy, 2009), including sidewalks (Ai & Tsai, 2016; Frackelton et al., 2013; Saha, Saugstad, Maddali, Zeng, Holland, Bower, Dash, Chen, Li, Hara, & Froehlich, 2019).

In this work, we address this data gap and take a step towards exploring the surface of our cities through CitySurfaces, a framework aimed at generating city-wide pavement material information by leveraging a collection of publicly available urban datasets. We combine active learning and computer vision-based segmentation model to locate, delineate, and classify sidewalk paving materials from street-level images. Our framework adopts a recent high-performing segmentation model (Tao, Sapra, & Catanzaro, 2020), which uses hierarchical multiscale attention combined with object-contextual representations. To tackle the challenges of high annotation costs associated with dense semantic label annotation, we make use of an iterative, multi-stage active learning approach, together with a previously acquired sidewalk inventory from Boston, which lists the dominant paving material for a given street segment. We demonstrate how the trained segmentation model can be extended with additional classes of materials with noticeably less effort, making it a versatile approach that can be used in cities with varying urban fabrics and paving materials. To show the generalization capabilities of CitySurfaces, we employ our framework in the segmentation of street-level images from four different cities: Brooklyn, Chicago, Washington DC, and Philadelphia, none of which were included in the training process. Fig. 1 highlights how different pavement materials are spatially distributed in three cities.

Our contributions can be summarized as follows:

- We present CitySurfaces, a deep-learning-based image segmentation framework for large-scale localization and classification of sidewalk paving materials.
- We adopt an active learning strategy to significantly reduce pixel-level annotation costs for training data generation, and yield increased segmentation accuracy.
- We conduct extensive experiments using street-level images from six different cities demonstrating that our model can be applied to cities with distinct urban fabrics, even outside of the domain of the training data.
- We make publicly available our model as well as the results of our material classification in the selected cities. The data can be accessed at: https://github.com/VIDA-NYU/city-surfaces.

This paper is organized as follows: Section 2 describes the main data sources of our framework; Section 3 describes the CitySurfaces framework; Section 4 summarizes our results; Section 5 highlights challenges and limitations; and Section 6 presents our conclusion.

2. Data description

Manually labeling the sidewalk materials in each image is a time-consuming task. Our proposed framework leverages a unique dataset that describes the material of sidewalks in Boston. We combine that data with the street-level images to create the training data for our semantic segmentation model. Next, we describe both data sources.

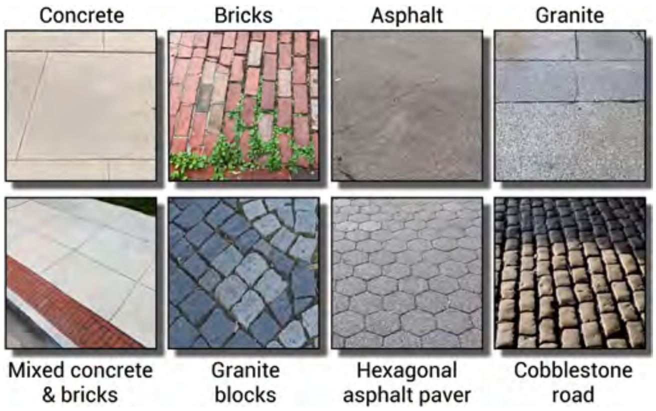

Fig. 2. The eight classes of surface materials used in our study. **Top**: Standard and prevalent materials. **Bottom**: Materials with distinct use.

2.1. *Boston sidewalk inventory*

The sidewalk inventory (Boston PWD, 2014) is part of the Boston Pedestrian Transportation Plan (Loutzenheiser, 2010) and describes sidewalk features, including geographic coordinates and paving materials collected via manual field visits. The material attribute describes the dominant surface material of each street segment (either concrete, brick, granite, a mix of concrete and brick, or asphalt). Fig. 2 illustrates patches of these five materials; the other three extra materials (granite block, cobblestone, hexagonal pavers) shown in the image were not recorded in the Boston dataset but were later manually added to our classes, as we will discuss in Section 3.3. We grouped the street segments by materials, using the geographic coordinates of the paving materials in the Boston inventory, and used it to assign an overall image class to the street-level images to guide the annotation process.

2.2. *Street-level images*

Street-level image usage in urban analysis has gained popularity with the introduction of Google Street View (GSV) (Anguelov et al., 2010) and Microsoft Street Slide (Kopf et al., 2010), services that provide panoramic images captured by specifically designed cameras mounted on a fleet of vehicles. These new data sources enable new questions and study designs for urban planning and design, urban sociology, and public health (Griew, Hillsdon, Foster, Coombes, Jones, & Wilkinson, 2013; Mooney et al., 2016; Yin, Cheng, Wang, & Shao, 2015). The GSV API can retrieve street-level images via geographic coordinates and allows users to adjust camera settings such as the heading, field of view (FoV), and pitch.

We use the OSMnx library (Boeing, 2017) to obtain the Boston street network and query the GSV API for street-level images at a fixed interval (5 m), excluding major highways and tunnels. We acquire the compass bearing of each street to set the camera heading to be perpendicular to the street, thus looking directly at left and right sidewalks. The pitch was set to 0° with an FoV of 80°. To create a more diverse training set, for 35% of the training data, we use different combinations of headings (pitch $\in [-10^\circ, -20^\circ]$, and FoV$\in [60^\circ, 70^\circ]$), to have sidewalk images taken at varying angles and perspectives. Fig. 3 illustrates sampled street segments in Boston, together with their image-level annotations. In order to train our framework, 3500 Boston images were obtained, and later 2000 images from New York City (NYC) were added to the pool of initially unannotated data. We excluded images with no sidewalks as well as those where more than 80% of the sidewalks were occluded. The final set had a total of 4300 images.

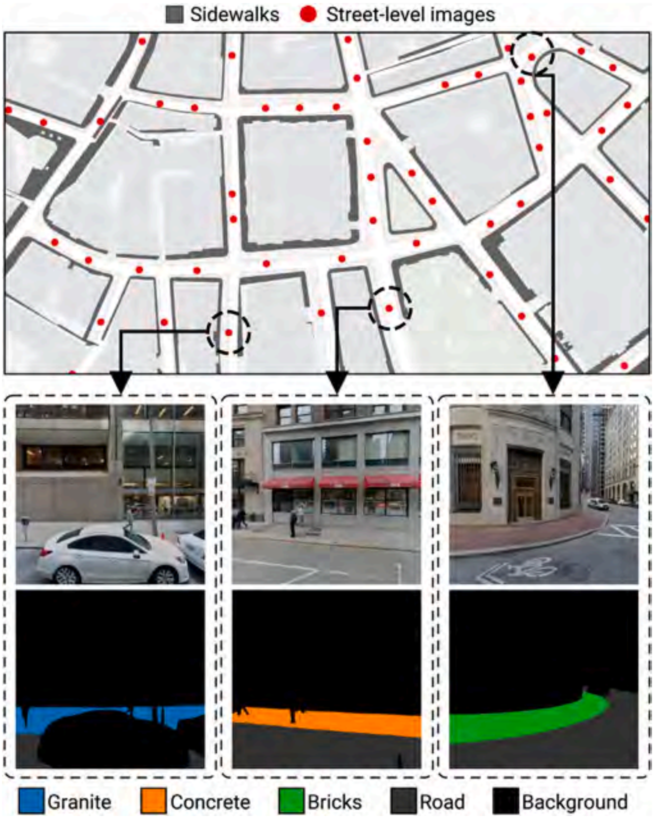

Fig. 3. Examples of sampled points in Boston to obtain street-level images. Three different sampling locations are highlighted and for each location, the street-level image as well as the prediction result of the model is depicted.

3. CitySurfaces

CitySurfaces adopts an active learning approach for the semantic segmentation of sidewalk paving materials. Using this framework, we aim to: (1) Train a model that can classify five different paving materials plus asphalt roads; (2) Extract information about sidewalk materials of a city for which no ground truth sidewalk inventory exists (e.g., NYC); and (3) Extend the model to classify additional classes of materials so that it can be applied to a more general set of cities.

Active learning aims at achieving high accuracy while minimizing the amount of required labeled data. The main hypothesis is, if we allow the model to choose the training data, it will perform better with fewer labeled instances (Settles, 2009). Through iteratively selecting the most informative or representative images to be labeled, fewer labeled instances are required to achieve similar performance, when compared to randomly selecting a large sample as training data and annotating all of it at once (Bloodgood & Vijay-Shanker, 2014; Huang, Jin, & Zhou, 2010).

In general, our multi-stage workflow is different from previous works in active learning for semantic segmentation in two important ways: First, our sample selection method is not fully automated; we use the uncertainty measure to filter the pool of unlabeled data in each stage, but we also use domain expertise for selecting a sample of images to be annotated and added to the training set in the next stage. Second, our query frequency is ten epochs (each epoch is a pass through all training data). The conventional approach in active learning is to select new samples (query) every iteration, which can work in cases where the cost of annotation is not high or in experimental studies that work with already annotated images to advance the field and develop new query algorithms, as is the case with most of the already published works in active learning for semantic segmentation, where they use

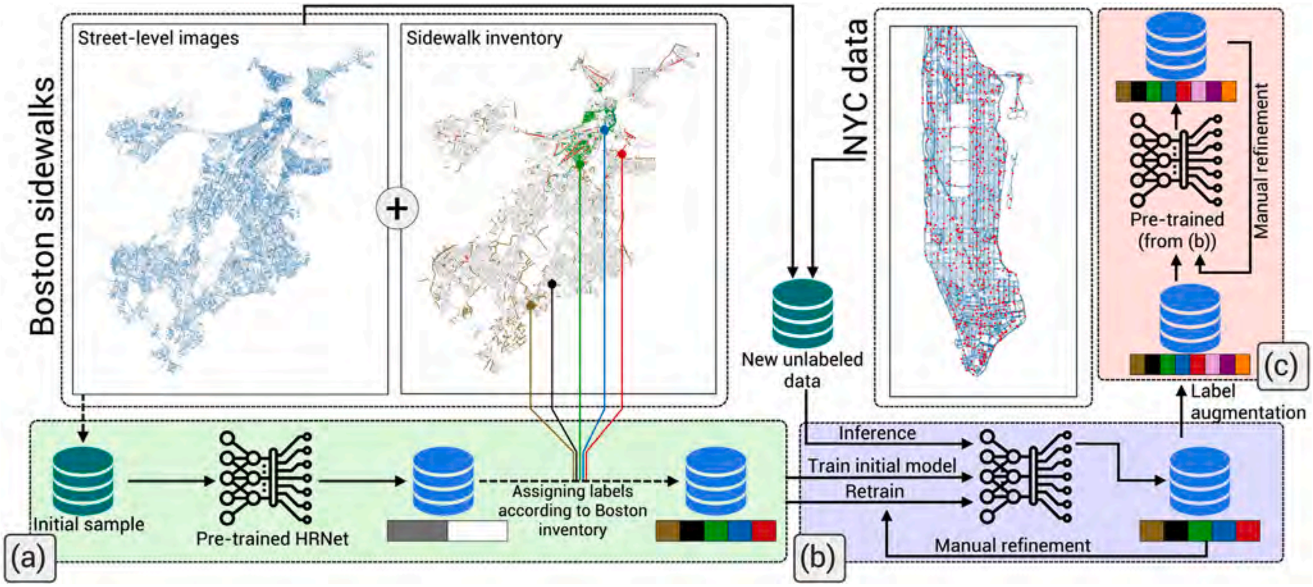

Fig. 4. CitySurfaces workflow. **Block (a)**: Creating the initial ground truth labels using the Boston sidewalk inventory and GSV images. A sample of unlabeled images is fed to a pre-trained HRNet, which outputs annotation labels containing two classes of interest: roads and sidewalks. The labels are manually refined to represent the five sidewalk paving classes, forming our ground truth set; **Block (b)**: Training the base model to classify five classes of surface materials, plus roads. The data from block (a) is used for the first stage of training. The model is then iteratively retrained for multiple stages on new samples. In each stage, the most representative and informative samples are chosen, and the annotations are manually refined and added to the training set to retrain the network; **Block (c)**: Introducing three new classes of materials. The pre-trained model from block (b) is retrained on the newly annotated image with three new classes. The final model can classify eight classes of different materials.

datasets such as Cityscapes (Cordts et al., 2016) or ADE20k (Zhou et al., 2017). However, since no annotated dataset exists for sidewalk materials, we have to annotate every new sample we choose during the training process, and it is impractical to annotate a new sample for every iteration (Kim et al., 2020). To overcome this, we adopt a multi-stage framework and annotate a new sample at the end of each stage, where each stage consists of ten epochs.

Our workflow has three major blocks as illustrated in Fig. 4: Block (a) creating initial training labels; Block (b) training a material segmentation model and; Block (c) extending the model to segment three additional classes from NYC standard materials. In this section, we first describe the different blocks of the workflow in detail, followed by a description of the semantic segmentation model. The training process and experiments were executed on 4 NVIDIA P100 GPUs with 12 GB of RAM each.

3.1. Block (a): Initial image annotation

To start the training process, we need a set of annotated images. To obtain the annotated data, we randomly sample 1000 images from a pool of unlabeled Boston street-level images and feed that sample into HRNet-W48 (Sun et al., 2019; Wang et al., 2020) model pre-trained on Cityscapes (Cordts et al., 2016) and get the initial segmentation results (Fig. 4(a)). The model outputs 19 classes from which we only keep roads and sidewalks. To generate an initial set of labeled data, we make use of the Boston Sidewalk Inventory (detailed in Section 2.1). We first query for the street segments of the images in our initial sample and modify the label to match the audited pavement from the inventory. Effectively, we are ensuring that, instead of having a general *sidewalk* class outputted by the pre-trained HRNet, our image set will have annotations according to the ground truth inventory data (e.g., concrete, bricks). We then manually refine them to account for the pre-trained model's prediction errors. In the initial training set, we restrict our sampling to images where the sidewalks mainly consist of a single material and eventually move to more complex material configurations in later stages. The final annotated images were split into 80% training and 20% validation to train the model in block (b).

3.2. Block (b): Model training on Boston and NYC

In the second block of the framework (Fig. 4(b)), we train an attention-based model (detailed in Section 3.4) using the labeled images from block (a). Our training step initially uses 800 images for training, and 200 images for validation, with a batch size of 8, SGD for the optimizer, momentum 0.9, weight decay $5e^{-4}$, and an initial learning rate of 0.002. We train the model in a multi-stage framework, where each stage consists of ten epochs. In each stage, we choose the epoch with the highest mIoU on the validation set. At the end of each stage, we make two decisions: (1) we select the best model considering all epochs of the current stage; and (2) we analyze the quantitative and qualitative results of the model to guide sampling the *new* addition to the training data. In particular, we analyze the confusion matrix, similarity matrix, as well as the top 10% of predictions with the highest mIoU and the top 20% of failures, obtained from the validation phase of the best epoch. The weights of the best model in the current stage are then used to initialize the model in the next stage with more training data. This restating scheme of SGD with the best solution of the previous stage is useful in increasing the chances of finding better solutions in the current stage.

To sample new images, we employ two strategies: (i) Uncertainty in predicting unlabeled images: We make use of the model's uncertainty estimations on unlabeled data and select the images that were most challenging for the model to predict; and (ii) Performance on validation set: By examining per-image IoU, uncertainty, and error rates of the images from failure and success cases together with confusion matrices, we construct a set of sample images to be used as inputs for finding similar unlabeled images. A more detailed explanation of these two techniques is provided in Appendix.

Following the sample selection strategies, we retrieve 300 unlabeled images, apply the current model on these new unlabeled images to generate a prediction, and then modify the predicted labels to add them to the overall training set, such that the segmentation model is trained on more samples of hard-to-segment images. To improve model generalization, in the third stage, we begin including images from Manhattan, which has a different urban fabric and more diverse forms and types of paving materials, in the pool of unlabeled data.

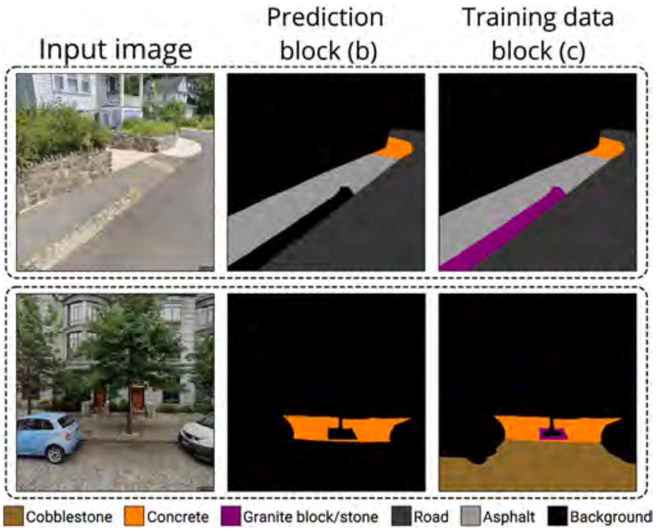

Fig. 5. Examples of how the annotation labels with additional classes were created from the output of the model in block (b) of our framework. The model trained in block (b) classified granite blocks and cobblestone as background, leaving smooth and clear boundaries, which helps to augment the labels with new classes during manual refinement and train a model that can classify eight different materials (block (c) of the framework).

Since no ground truth data exists for Manhattan, to create the ground truth label, we need to have a model with reliable performance to create the base annotation. We chose the third stage since the model reached a reliable performance (83% mIoU) in detecting the main classes, and outputs had clear borders compared to the other two stages. The selected images from Manhattan were fed to the model, and the results were corrected and refined using feedback from the domain expert and added to the training dataset. The segmentation model is then trained on the combined set of the initial and newly annotated data (1100 images), initialized with the weight from the best epoch of the previous stage. This procedure is iterated for five stages (at which point we observe no further notable improvements). The model at the final stage was trained on 2500 images (Fig. 4(b)), and achieved 88.6% mIoU on the held-out test set.

3.3. Block (c): Including additional materials from NYC

Once the model in block (b) attains sufficiently accurate segmentation performance, we extend it by adding three additional classes (Fig. 4(c)). The three additional classes are granite blocks, hexagonal pavers, and cobblestone. These materials are standard sidewalk materials in the NYC street design manual (NYC DOT, 2020). While granite blocks and cobblestones were also observed in Boston, they were not included in the Boston sidewalk inventory. Since the original model in block (b) was not trained to detect these materials, they are initially either classified as background (mostly granite blocks and cobblestones) or misclassified (mostly hexagonal pavers) as other visually similar materials. To collect street-view images that have these new materials, we follow the NYC and Boston street design manuals (NYC DOT, 2020; Thomas M. Menino, 2013) to filter unlabeled data from the locations in which these materials can be found. For example, hexagonal pavers (NYC only) are typically used on sidewalks adjacent to parks and open spaces, and cobblestones are used in historic districts.

We select a total of 800 images that contain these new classes to be iteratively sampled for training, 150 additional images for the validation set, and 200 images for the held-out test set. Annotating the new image set consumed fewer resources as compared to the initial annotations since smooth model predictions typically leave clear boundaries, which only needed to be assigned the appropriate label (see Fig. 5). The newly generated set of labels was used to train the model by initializing the architecture with model weights in block (b) and only replacing the final softmax layer instead, to produce ten output channels (corresponding to eight paving materials, plus the road, and background). At the end of each stage, we select a new sample of unlabeled images following the same process explained in Section 3.2, run them through the model, obtain segmentation predictions, refine the results, and retrain the model. In total, 726 additional images were added to the training set, and in the final stage, the model was trained on 3226 images (2500 from block (b) + 726). We halt the training in stage 3 after 30 epochs, and test the model on the held-out test set (Fig. 4(c)). Fig. 6 shows the confusion matrices for all three stages of our extended model, illustrating model performance as a function of the amount of training data. These matrices were also used in part to guide the sampling of images to annotate.

Using the described method, model performance increases from 74.3% mIoU to 88.6% for the base model (block (b)) and to 90.5% in the extended model (block (c)), with the manual refinement time decreasing from 25 to 4 min per image. Fig. 7 depicts the evolution of the segmentation results of block (c) through the active learning stages. The model outputs more refined boundaries and significantly less noise in later stages; thus, significantly less time is needed to modify the newly annotated data as the stages go on. In each stage, the model is initialized with the weights from the previous stage.

3.4. Semantic segmentation model

For the semantic segmentation task (blocks (b) and (c)), we adopt the Hierarchical Multi-Scale Attention (Tao et al., 2020) and fine-tune the parameters on our dataset. To train the model, following Zhu et al. (2019), we employ class uniform sampling in the data loader, which chooses equal samples for each class for handling the class imbalance, since some classes like road and background are almost present in all images, whereas classes like cobblestone and hexagonal pavers are not that prevalent. The Region Mutual Information (RMI) loss (Zhao, Wang, Yang, & Cai, 2019) was employed as the primary loss function. RMI takes the relationship between pixels into account and uses the neighboring pixels around each pixel to represent it instead of only relying on single pixels to calculate the loss. We run different experiments with and without the RMI loss function for the main segmentation head. In the absence of RMI, standard cross-entropy loss was used instead. The model under the same setting, but without RMI loss, performed slightly worse (89.84) compared to the one where RMI loss was used (90.51). Fig. 8 presents an overview of the architecture. Next, we describe the network's architecture in more detail.

3.4.1. Backbone

We chose HRNet-OCR (Yuan, Chen, & Wang, 2019) as the backbone. The network comprises HRNet-W48 (Sun et al., 2019; Wang et al., 2020) and adds Object-Contextual Representations (Yuan et al., 2019) to further augment the representation extracted by the HRNet. The final representation from HRNet-W48 works as the input to the OCR module, which computes the weighted aggregation of all the object region representations to augment the representation of each pixel. The weights are calculated based on the relations between pixels and object regions. The augmented representations are the input for the attention model described next.

3.4.2. Attention model

The model is mainly based on Share-Net (Chen, Yang, Wang, Xu, & Yuille, 2016). Suppose an input image is resized to several scales, i.e., $s \in \{1, \ldots, S\}$. Each scale is passed through the backbone part (HRNet-W48+OCR), and we can get the output feature $f_{i,c}^{s}$. For the feature, $c \in \{1, \ldots, C\}$ (C is the number of classes of interest, and i ranges over all the spatial positions). As shown in Fig. 8, the features then go through two heads, one for attention generation and the

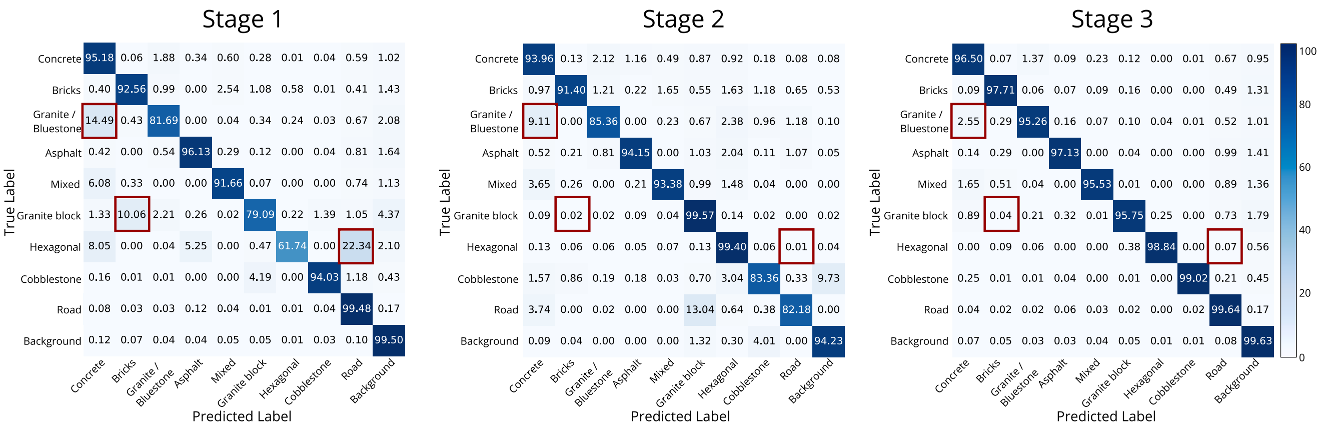

Fig. 6. Confusion matrices for the three stages of the extended model. These results guided sample selection and signaled which type of images should be included in the training data for the next stage. Notice the improvement of the predictions for hexagonal pavers, granite block, and granite/bluestone (highlighted in red).

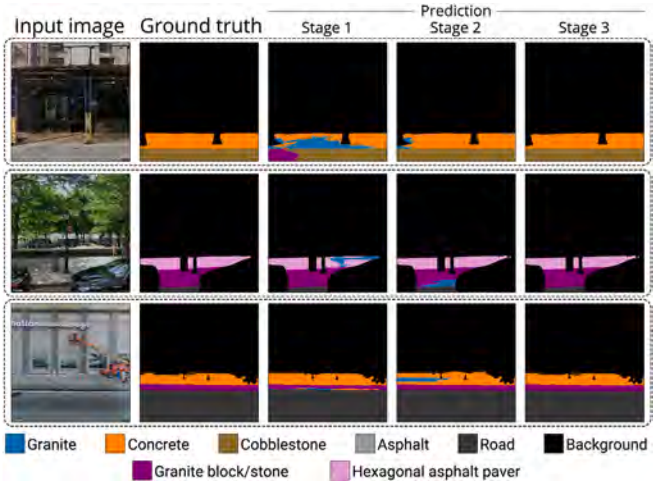

Fig. 7. Evolution of the block (c) extended model's inference through different training stages.

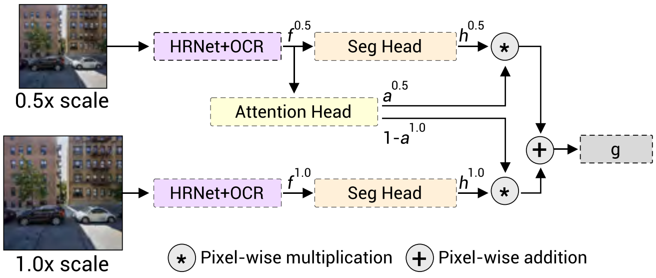

Fig. 8. The general architecture of the hierarchical multi-scale attention (HMSA) based semantic segmentation method (Tao et al., 2020). The inputs are images from two scales. The network learns the relative attention between scales and hierarchically applies the learned attention to combine the results from two segmentation heads and make a prediction.

other for segmentation. The features $f^{s}_{i,c}$ are resized for different scales to have the same resolution (with respect to the finest scale) using bilinear interpolation before passing the model heads. For the attention head, we generate the learned weights for $f^{s}_{i,c}$ which is represented by $a^{s}_{i,c}$. This weight is integrated into the initial output $h^{s}_{i,c}$ from the segmentation head, and we have:

$$g^{s}_{i,c} = a^{s}_{i,c} * h^{s}_{i,c} \quad (1)$$

in which $g^{s}_{i,c}$ is the final output score map for scale s, and $*$ here represents the pixel-wise multiplication.

In the model, the combination of score maps is similar to Tao et al. (2020) to make the flexible scales during inference time possible and improve the training efficiency. During the training, we only need to train with two adjacent scales (as shown in Fig. 8). During testing, weights for the network are shared for each adjacent scale pair.

To be more specific, suppose the two selected adjacent scales are $1x$ and $0.5x$ (the final selected scales during training in the model are $0.5x$, $1x$, and $2x$) to obtain the pair of scaled images for the model input. For inference, we can hierarchically and repeatedly use the learned attention to combine N scales of predictions together. Precedence is given to lower scales since they have a more global context and can choose where predictions need to be refined by higher scale predictions. The final combination principle for these adjacent scales is defined as:

$$g_{i,c} = a^{0}_{i,c}.5 * h^{0}_{i,c}.5 + (1 - a^{0}_{i,c}.5) * h^{1}_{i,c} \quad (2)$$

The hierarchical mechanism used in the model coupled with the powerful HRNet-OCR backbone provides a robust architecture for the challenging task of material classification *in the wild.*

4. Results

In this section, we present the results of applying our trained model on the held-out test set. We do not rely on pixel-level accuracy in evaluating the model since sidewalks comprise a relatively small portion of each image, while road and background can occupy more than 70% of most images, resulting in a significant class imbalance. This class imbalance creates an arbitrary high pixel-level accuracy, which is not a fair representation of the model's performance.

4.1. General evaluation metrics

Table 1 presents class-level evaluation metrics, the mean Jaccard index (IoU), precision, and recall for the final model. The model outputs ten classes in total, seven classes of sidewalk pavings, one extra class of street pavings (cobblestone), plus road and background. Excluding road and background, the model achieved 88.37% accuracy, with hexagonal asphalt pavers and asphalt sidewalks having the highest accuracy measures. Overall, half of the pavement classes have IoU above 90%. Concrete, the most prevalent and versatile material, can be classified with 88.7 accuracy. A robust result considering the high within-class variation (i.e., it comes in various colors and textures). Granite/bluestone and granite block have the lowest accuracy (81.09 and 82.92 respectively). This can be partially explained by their visual similarity to dark concrete (or wet concrete), potentially leading to more false positive predictions.

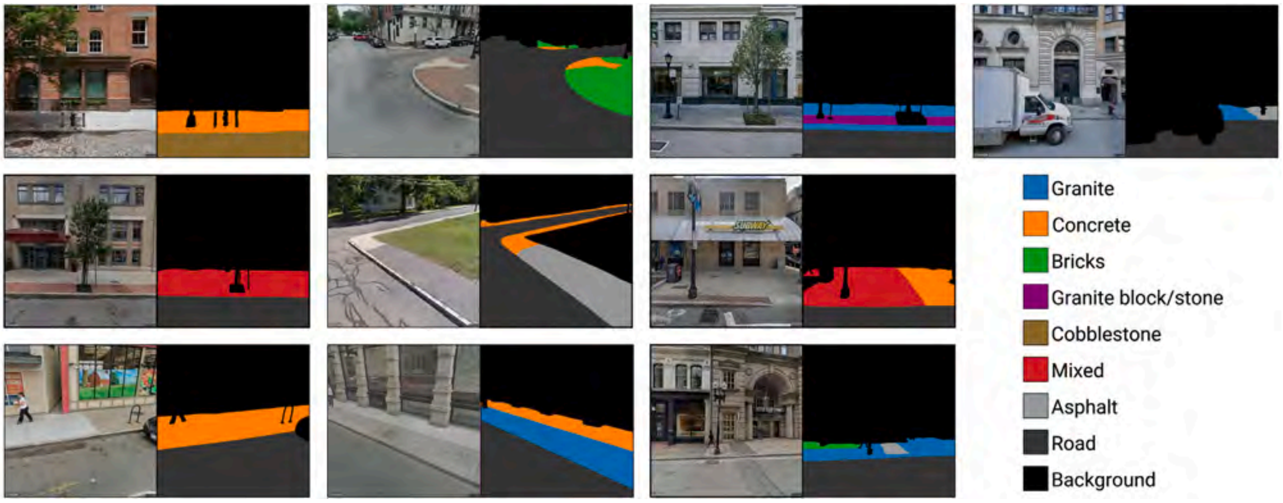

Fig. 9. Predictions of the model on the held-out test set. Fine details and boundaries of objects like poles, plants, wooden sticks, and fire hydrants are very precisely predicted. The model also segmented curb cuts (line 1 - column 2), different instances of the same material (3-1 and 3-3), and visually similar materials of different classes (1-4).

Table 1
Evaluation metrics on the held-out test set.

Label	IoU	Precision	Recall
Concrete	88.69	0.95	0.93
Brick	91.79	0.95	0.96
Granite/Bluestone	81.09	0.85	0.95
Asphalt	92.58	0.96	0.97
Mixed	86.11	0.93	0.93
Granite block/Stone	82.92	0.94	0.88
Hexagonal asphalt paver	92.81	0.98	0.95
Cobblestone	90.95	0.94	0.96
Road	99.01	0.99	1
Background	99.16	1	1
mIoU		90.51	
mIoU (eight main classes)		88.37	

Table 2
Evaluation metrics on samples from the selected cities (outside of training domain).

City	mIoU	Mean per-segment accuracy
Brooklyn	86.12	87.09
Chicago	84.31	86.52
Washington DC	82.61	84.27
Philadelphia	82.81	83.46

Fig. 9 illustrates some examples of the model's prediction, highlighting its performance in detecting boundaries between fine objects, like poles and plants, even in shadowed scenes (line 1 - column 1, 1-3, 2-1). The model can also detect curb ramps in most scenes, even though it was not specifically trained with such a goal (1-1 and 2-2). Fig. 9 (1-2) shows an example in which the model accurately classified a sidewalk segment with patches of different materials. We can also see the model performance in distinguishing between visually similar materials (1-4, 3-2), as well as different variation of the same material such as (3-1) where two visually distinct concrete slabs are classified correctly.

4.2. Evaluating the generalization capabilities of CitySurfaces

To demonstrate the generalization capabilities of CitySurfaces, we tested the performance of our approach on samples from Chicago, Washington DC, Philadelphia, and Brooklyn, which were not part of the training data. We randomly sampled 200 street segments from each city, and obtained their corresponding street-view images, at every five meters of each segment, from the left and right sides of the sidewalks. After data cleaning and pre-processing, we were left with roughly 600 images per city; these images were annotated using the model in block (b), then *manually* checked and refined to create the test set. Table 2 shows the results of applying CitySurfaces on these test sets. We report mIoU and mean per-segment accuracy. Mean per-segment is a simple and practical metric that measures whether the model correctly detected the dominant materials in each street segment and report the average accuracy over all images in the test set. All tested cities had an accuracy greater than 82%. Brooklyn achieved the highest accuracy, since the borough's paving materials follow the same street design regulation as Manhattan, which was part of the training data.

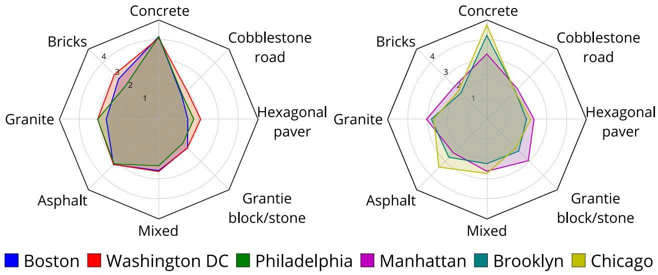

Fig. 10. Comparison of the distribution of detected materials in six different cities. The star plots show the log of the number of sidewalk segments identified as having a given material.

CitySurfaces enables generating city-wide sidewalk material datasets, as illustrated in Fig. 9. This allows us to compare the distribution of different paving materials in various cities. Fig. 10 shows the result of this comparison. We can see that Manhattan and Washington DC use more diverse and balanced material types. Concrete is the dominant material in all of the cities. Chicago has the highest number of asphalt sidewalks among the selected cities; Boston, Washington DC, and Philadelphia have a similar number of asphalt sidewalks, which come second to Chicago. Asphalt sidewalks are mainly used in suburban neighborhoods; that is why dense urban areas like Manhattan and Brooklyn have the lowest number of sidewalks paved with asphalt. Another interesting observation is the higher usage of granite/bluestone in Manhattan compared to Brooklyn, two boroughs of the same city. Granite is considered an expensive and decorative material, used mainly in commercial streets or historic neighborhoods, which signals Manhattan's higher land value and income level, since maintenance and installation of decorative pavings are the owner's responsibility.

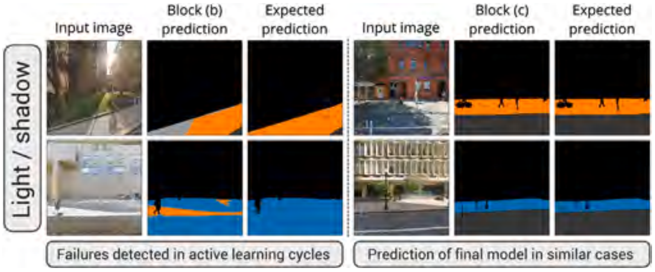

Fig. 11. **Left**: Exposure to direct sunlight changed the appearance of colors and texture of the paving material, **Left top**: Part of a concrete sidewalk under the shadow was classified as asphalt. **Left bottom**: Part of a granite surface under direct sunlight was classified as concrete. **Right**: The correct predictions of the final model in the same settings.

5. Discussion

The specific characteristics of computing the spatial distribution of sidewalk pavement materials require experts to oversee the performance of the model and ensure that the network is correctly classifying the pavement materials. Through active learning process, we identified certain elements of the urban scenes that can create higher prediction confusion and lead to misclassification. Two main categories of patterns repeatedly observed among the failure cases were shadow/light contrasts (Fig. 11) and distinct objects such as metal gratings and plant pits that resemble brick from a distance (Fig. 12). The texture and color of different materials can appear different under shadow or extreme light, showing a higher resemblance to another material. For instance, under the shadow, concrete is classified as asphalt (Fig. 11 - left top). Moreover, some patterns or objects can look similar to certain materials. For example, the model initially classified certain plant pits (Fig. 12 - left top) or brownish metal covers (Fig. 12 - left bottom) as bricks alongside the concrete pavement and would incorrectly predict mixed materials for that part of the sidewalk, or even small pieces of broken concrete or granite were classified as cobblestones (Fig. 12 - left middle). Adding more images with these patterns to the training data improved the model's performance in the next stage. Some examples of the correct predictions for similar patterns are shown on the right side of Fig. 12. The active learning strategy significantly helped with choosing the right data at each stage. Having an expert in the loop to review the results in each stage enabled identifying specific patterns that were not evident by merely analyzing the quantitative metrics of the model.

5.1. Challenges

One of the key challenges of this study was handling different textures of the same object (sidewalk). Objects have defined boundaries that are easier to classify (Jain & Gruteser, 2018). However, similar textures can appear on multiple objects. For instance, red bricks are used in both building facades and sidewalk pavings (although different types of bricks are used for each purpose, they possess very close visual characteristics). Our goal is to have a model that can detect *sidewalks* of certain materials from street-view images.

Another challenging aspect of this task is the high degree of within-class variation and between-class similarities. For instance, NYC designated five different types of concrete as standard materials for sidewalk pavings, while Boston uses three different types of concrete. Each of these types has distinct visual features that, in some cases, can resemble materials of other classes, which pose further challenges to the classification task. Distinguishing between dark concrete and bluestone in some cases is very difficult, even for humans. When wet, some concretes with aggregates can look very similar to granite, and under the shadow, asphalt and worn-off concrete can look very similar. Having a model that can accurately handle the within-class variability with between-class similarity calls for smartly selected training datasets with a good distribution of different classes as well as multiple variants of the same material under different conditions.

5.2. Limitations

Even though CitySurfaces can provide city-scale sidewalk material classification, some challenges remain unaddressed. For instance, in the absence of proper sidewalk network data, it can be challenging to map the materials to their corresponding locations accurately. The maps in Fig. 1 are based on the road centerlines where GSV cars traveled to capture images, depicting the dominant materials for each street segment by taking an average over the materials observed in each image from both the left and right sides of the street. However, knowing the exact location of certain materials is critical for urban designers, planners, and those working with safety and ease of walk for people with special needs. Although our model produces this result at a highly fine level, we cannot depict this variety in detail without proper sidewalk network data. Having separate maps for left and right sidewalks can be one solution, but the intersections where more than one street is captured pose a challenge for assigning the correct materials to each segment.

Also, street-level images have some inherent limitations. Since the images are taken by cars moving alongside streets, in many instances, specifically in dense urban areas, the cars parked on the sides blocked the sidewalk view, as shown in the left image of Fig. 3. The issue can be mitigated to some extent by adjusting the heading and pitch of the camera, but that solution fails in images with large vehicles like trucks, or when the car with mounted cameras is too close to the sidewalks.

6. Conclusion

We present CitySurfaces, a scalable, low-cost approach towards the automatic computation of the spatial distribution of pavement materials at the sidewalk segment level. Our model can detect a diverse range of materials, which to our knowledge, were not covered by any existing dataset. For instance, hexagonal pavers or granite blocks were not reported in any sidewalk inventories reviewed in this study. CitySurfaces produces accurate segmentation considering multiple cities both within and outside the domain of the training data, demonstrating generalization capabilities across varying urban fabrics. CitySurfaces can detect, delineate, and classify eight standard surface materials used throughout most US cities. As shown in Section 3.3, the framework can be extended to include additional surface materials with less effort than building a city-specific model from scratch, which makes it possible for almost any city or government agency that has spatially dense street-level image data, to create a similar dataset. Moreover, since we have covered the standard materials, such as concrete, asphalt, granite/bluestone, and brick, the model can be applied to a wide range of cities without any further annotation effort or with substantially less effort using our pre-trained model. The models and generated datasets for the selected cities are publicly available in a GitHub repository.

This work has addressed some challenges in data annotation and accurate classification of different materials with high between-class similarities and within-class variation. The active learning framework utilized in this study helped reduce the annotation costs by choosing the most informative set of data to be annotated and incrementally decreasing the manual modification time. By offering the first comprehensive dataset of sidewalk surface materials at the city scale, this study goes beyond reporting the dominant material of each segment and provides information on the percentage distribution of all detected materials per sidewalk segment. The material classes in this study were selected based on the standard surface materials listed by Boston sidewalk inventory (Boston PWD, 2014), to use it as our baseline ground truth. That list is not extensive and does not distinguish between various types of the same class of material, such as concrete. However,

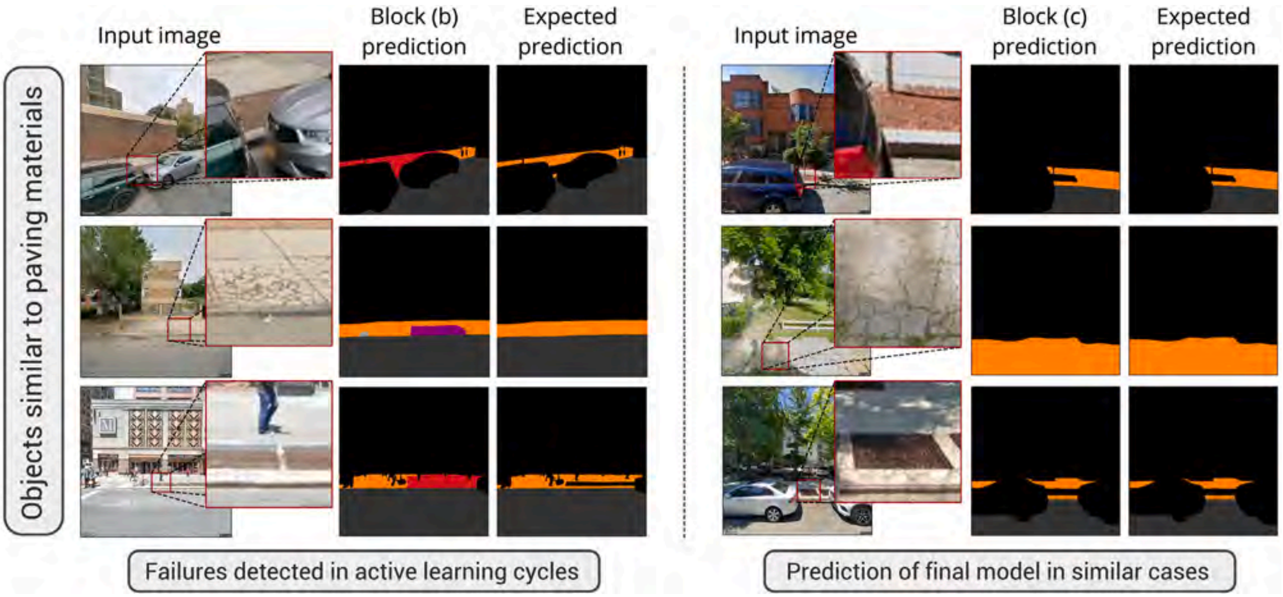
Objects similar to paving materials
Input image
Block (b) prediction
Expected prediction
Failures detected in active learning cycles
Input image
Block (c) prediction
Expected prediction
Prediction of final model in similar cases

Fig. 12. Objects with patterns similar to different materials. **Left**: Classifying failures caused by different patterns. **Left top**: Concrete alongside a furnishing zone was misclassified as mixed class since plant pit was detected as bricks. **Left middle**: Broken concretes were misclassified as granite blocks. **Left bottom**: Concrete was misclassified as mixed class due to the presence of brownish metal covers. **Right**: Correct prediction of the model for the similar pattern in the final cycle of active learning.

for some more in-depth analysis, such as measuring UHI, we may need to classify the materials differently, and distinguish between different variations of the same material within one class. For instance, reflective granite and dark matte bluestone should have two distinct classes, same goes with the dark and light concretes since they have distinctively different albedo values. The CitySurfaces framework can be easily extended to detect more classes of materials as illustrated with the Manhattan example in Section 3.3, given the availability of the images corresponding to each class of interest to create the initial ground-truth set. In future work, we plan to take these differences into account and combine the generated data with shadow accumulation (Miranda et al., 2019) to generate a city-scale UHI map.

To facilitate designing automated audit tools, we are going to extend our model to detect surface problems such as potholes, significant breakage, and obstacles on pedestrian paths for accessibility analysis (Miranda et al., 2020). We also aim to address the walkability and active design of sidewalks by developing a model to detect relevant features of the sidewalk's wall plane and furnishing zone, such as window-to-wall ratio. As another line for our future work, we would like to explore automated sample selection procedures and self-supervised learning techniques and tailor them to sidewalk and pedestrian facility analysis. We chose a simple (yet effective) uncertainty measure and coupled it with the analysis of the model's performance on the validation set and used expert's feedback to refine the annotations and check whether the model is predicting correctly since, on many instances, it is difficult to distinguish between visually similar materials.

CRediT authorship contribution statement

Maryam Hosseini: Conceptualization, Methodology, Data Curation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. **Fabio Miranda:** Conceptualization, Visualization, Data Curation, Writing – original draft, Writing – review & editing. **Jianzhe Lin:** Methodology, Writing – original draft. **Claudio T. Silva:** Supervision, Writing – review & editing, Funding acquisition.

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.

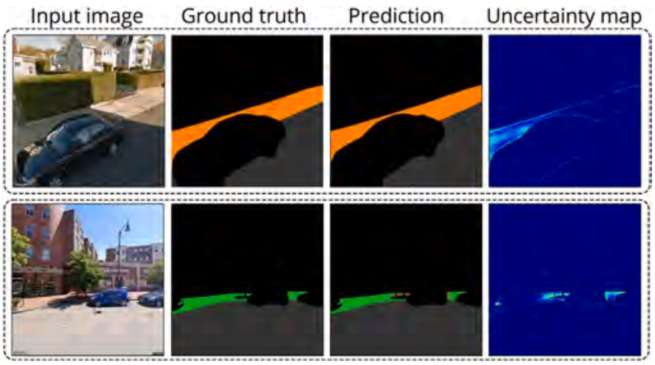

Fig. 13. Two different scenarios of using the model's output and uncertainty map in sample selection. The warmer colors in the uncertainty map represent areas where the model was less confident in its prediction. **Top**: The model correctly predicted the class in a previously identified challenging setting (shadow) but was less certain in predicting the shadowed areas. **Bottom**: The model classified the parts in shadow as concrete alongside brick and outputted mixed class for that region. The uncertainty map shows that the model was least certain in its prediction for that area.

Acknowledgments

We would like to thank our colleagues at New York University for their help in this research. This work was supported in part by: the Moore-Sloan Data Science Environment at NYU; C2SMART at NYU; NASA; NSF awards CNS-1229185, CCF-1533564, CNS-1544753, CNS-1730396, CNS-1828576, CNS-1626098; and the NVIDIA NVAIL at NYU. Claudio T. Silva is partially supported by the DARPA D3M program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA. We also thank NVIDIA Corporation for donating GPUs used in this research.

Appendix. Sampling strategies

(i) Uncertainty in predicting unlabeled images. Uncertainty sampling is one of the most frequently used query methods to select a new sample

of training data in active learning (Settles, 2009). To measure the uncertainty, we use softmax probability, which has been commonly used in active learning as a strategy for choosing the next training sample (Settles, 2009). We use the outputs of the softmax layer as part of the sampling strategy, which can partly reveal the most challenging instances for the model to predict. We apply multi-class uncertainty sampling known as margin sampling (MS) (Scheffer, Decomain, & Wrobel, 2001), which calculates the difference between the two highest prediction probabilities on softmax to produce uncertainty maps. The smallest margin in each map is then chosen as the image-level uncertainty. The MS measure is defined as:

$$x^*_{MS} = argmin_x P_\theta(\hat{y}_1|x) - P_\theta(\hat{y}_2|x) \tag{A.1}$$

where $\hat{y}_1$ and $\hat{y}_2$ are the class labels for pixel x, with the first and second highest probability, respectively, under the model θ. The lowest margin gives us the highest uncertainty, which is used as an image-level uncertainty measure.

To select new samples, we feed the pool of unlabeled images to our network, obtain the segmentation and calculate image-level uncertainty to select images with the highest uncertainty. We start by selecting 10% of the images using this strategy. As the training proceeds, we increase the share of images selected through this strategy at each stage by 10%.

(ii) Performance on validation set. Since softmax probabilities do not necessarily represent the true *correctness* likelihood, a problem known as "confidence calibration" (Guo, Pleiss, Sun, & Weinberger, 2017), we need other strategies as well to select an informative sample for the model. To this end, at each stage, we examine the performance of the best epoch on the validation set and select 10% of the best predictions and 20% of the top failures. Images from failure and success cases are then clustered using K-means (Cover & Hart, 1967; Fix, 1985) with the Euclidean distance to investigate potential common patterns in each group. In each cluster, we rank images based on the average IoU of all classes, excluding road and background. We then select images with the highest error rate. The error rate is defined as the sums of false positive and false negative predictions of the model in each image. Aside from the described method, we examine the clusters of images to detect common error-causing patterns. Fig. 13 (bottom row) depicts a brick sidewalk that the initial model incorrectly segmented the part next to shadowed regions as the "mixed" class. Its associated uncertainty map reveals prediction difficulty near the edge of the car and the plant pit, which are incorrectly classified as mixed. Uncertainty maps of the success cases are examined to find regions where the model is least confident while making a correct prediction. Fig. 13 highlights a set of uncertainty maps. After we find the most error-prone images, we use them to find similar unlabeled images. We extract their feature maps using the backbone HRNet-W48 (Sun et al., 2019; Wang et al., 2020) (more details in Section 3.4.1) and employ cosine similarity distance to retrieve similar images from the pool of unlabeled data.

References

Agathangelidis, I., Cartalis, C., & Santamouris, M. (2020). Urban morphological controls on surface thermal dynamics: A comparative assessment of major European cities with a focus on athens, Greece. *Climate*, *8*(11), 131.

Aghaabbasi, M., Moeinaddini, M., Shah, M. Z., Asadi-Shekari, Z., & Kermani, M. A. (2018). Evaluating the capability of walkability audit tools for assessing sidewalks. *Sustainable Cities and Society*, *37*, 475–484.

Ai, C., & Tsai, Y. (2016). Automated sidewalk assessment method for Americans with disabilities act compliance using three-dimensional mobile lidar. *Transportation Research Record: Journal of the Transportation Research Board*, (2542), 25–32.

Akbari, H., Menon, S., & Rosenfeld, A. (2009). Global cooling: increasing world-wide urban albedos to offset CO 2. *Climatic Change*, *94*(3), 275–286.

Akbari, H., & Rose, L. S. (2008). Urban surfaces and heat island mitigation potentials. *Journal of the Human-Environment System*, *11*(2), 85–101.

Amati, M., & Taylor, L. (2010). From green belts to green infrastructure. *Planning Practice & Research*, *25*(2), 143–155.

Anguelov, D., Dulong, C., Filip, D., Frueh, C., Lafon, S., Lyon, R., et al. (2010). Google street view: Capturing the world at street level. *Computer*, *43*(6), 32–38.

Arnold, C. L., Jr., & Gibbons, C. J. (1996). Impervious surface coverage: the emergence of a key environmental indicator. *Journal of the American planning Association*, *62*(2), 243–258.

Bell, C. D., Tague, C. L., & McMillan, S. K. (2019). Modeling runoff and nitrogen loads from a watershed at different levels of impervious surface coverage and connectivity to storm water control measures. *Water Resources Research*, *55*(4), 2690–2707.

Bloodgood, M., & Vijay-Shanker, K. (2014). A method for stopping active learning based on stabilizing predictions and the need for user-adjustable stopping. arXiv preprint arXiv:1409.5165.

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

Boston PWD (2014). Boston sidewalk inventory. URL: https://data.boston.gov/dataset/sidewalk-inventory.

Chen, L.-C., Yang, Y., Wang, J., Xu, W., & Yuille, A. L. (2016). Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3640–3649).

Chen, X., & Zhang, Y. (2017). Impacts of urban surface characteristics on spatiotemporal pattern of land surface temperature in kunming of China. *Sustainable Cities and Society*, *32*, 87–99.

Chippendale, T., & Boltz, M. (2015). The neighborhood environment: perceived fall risk, resources, and strategies for fall prevention. *The Gerontologist*, *55*(4), 575–583.

Chithra, S., Nair, M. H., Amarnath, A., & Anjana, N. (2015). Impacts of impervious surfaces on the environment. *International Journal of Engineering Science Invention*, *4*(5), 27–31.

Clifton, K. J., Smith, A. D. L., & Rodriguez, D. (2007). The development and testing of an audit for the pedestrian environment. *Landscape and Urban Planning*, *80*(1), 95–110.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., et al. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3213–3223).

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27.

Deitz, S., Lobben, A., & Alferez, A. (2021). Squeaky wheels: Missing data, disability, and power in the smart city. *Big Data & Society*, *8*(2).

Du, H., Cai, W., Xu, Y., Wang, Z., Wang, Y., & Cai, Y. (2017). Quantifying the cool island effects of urban green spaces using remote sensing data. *Urban Forestry & Urban Greening*, *27*, 24–31.

Estoque, R. C., Murayama, Y., & Myint, S. W. (2017). Effects of landscape composition and pattern on land surface temperature: An urban heat island study in the megacities of southeast Asia. *Science of the Total Environment*, *577*, 349–359.

Ewing, R., & Handy, S. (2009). Measuring the unmeasurable: Urban design qualities related to walkability. *Journal of Urban Design*, *14*(1), 65–84.

Fix, E. (1985). *Discriminatory analysis: Nonparametric discrimination, consistency properties, Vol. 1*. USAF school of Aviation Medicine.

Frackelton, A., Grossman, A., Palinginis, E., Castrillon, F., Elango, V., & Guensler, R. (2013). Measuring walkability: Development of an automated sidewalk quality assessment tool. *Suburban Sustainability*, *1*(1), 4.

Glaeser, E. L., Kominers, S. D., Luca, M., & Naik, N. (2018). Big data and big cities: The promises and limitations of improved measures of urban life. *Economic Inquiry*, *56*(1), 114–137.

Griew, P., Hillsdon, M., Foster, C., Coombes, E., Jones, A., & Wilkinson, P. (2013). Developing and testing a street audit tool using google street view to measure environmental supportiveness for physical activity. *International Journal of Behavioral Nutrition and Physical Activity*, *10*(1), 1–7.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International Conference on Machine Learning* (pp. 1321–1330). PMLR.

Huang, S.-J., Jin, R., & Zhou, Z.-H. (2010). Active learning by querying informative and representative examples. *Advances in Neural Information Processing Systems*, *23*, 892–900.

Jain, S., & Gruteser, M. (2018). Recognizing textures with mobile cameras for pedestrian safety applications. *IEEE Transactions on Mobile Computing*, *18*(8), 1911–1923.

Joshi, P., Leitão, J. P., Maurer, M., & Bach, P. M. (2021). Not all SuDS are created equal: Impact of different approaches on combined sewer overflows. *Water Research*, *191*, Article 116780.

Kim, T., Hwang, I., Lee, H., Kim, H., Choi, W.-S., & Zhang, B.-T. (2020). Message passing adaptive resonance theory for online active semi-supervised learning. arXiv preprint arXiv:2012.01227.

Kopf, J., Chen, B., Szeliski, R., & Cohen, M. (2010). Street slide: browsing street level imagery. *29*, (4), (pp. 96:1–96:8). ACM,

Lay, M., Metcalf, J., & Sharp, K. (2020). *Paving our ways: a history of the world's roads and pavements*. CRC Press.

Li, X., Zhou, W., & Ouyang, Z. (2013). Relationship between land surface temperature and spatial pattern of greenspace: What are the effects of spatial resolution? *Landscape and Urban Planning*, *114*, 1–8.

Loutzenheiser, F. (2010). Boston region's pedestrian transportation plan. URL: https://www.mapc.org/wp-content/uploads/2017/11/PedPlanFullReport.pdf.

Miranda, F., Doraiswamy, H., Lage, M., Wilson, L., Hsieh, M., & Silva, C. T. (2019). Shadow accrual maps: Efficient accumulation of city-scale shadows over time. *IEEE Transactions on Visualization and Computer Graphics*, *25*(3), 1559–1574.

Miranda, F., Hosseini, M., Lage, M., Doraiswamy, H., Dove, G., & Silva, C. T. (2020). Urban mosaic: Visual exploration of streetscapes using large-scale image data. In *CHI'20, Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–15). Association for Computing Machinery.

Mooney, S. J., DiMaggio, C. J., Lovasi, G. S., Neckerman, K. M., Bader, M. D., Teitler, J. O., et al. (2016). Use of google street view to assess environmental contributions to pedestrian injury. *American Journal of Public Health*, *106*(3), 462–469.

Muench, S. T., Anderson, J., & Bevan, T. (2010). Greenroads: A sustainability rating system for roadways.. *International Journal of Pavement Research & Technology*, *3*(5).

Nwakaire, C. M., Onn, C. C., Yap, S. P., Yuen, C. W., & Onodagu, P. D. (2020). Urban heat island studies with emphasis on urban pavements; a review. *Sustainable Cities and Society*, Article 102476.

NYC DOT (2020). Street design manual.

Oke, T. R. (1982). The energetic basis of the urban heat island. *Quarterly Journal of the Royal Meteorological Society*, *108*(455), 1–24.

Saha, M., Saugstad, M., Maddali, H. T., Zeng, A., Holland, R., Bower, S., et al. (2019). Project sidewalk: a web-based crowdsourcing tool for collecting sidewalk accessibility data at scale. In *CHI '19, Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). Association for Computing Machinery.

Santamouris, M. (2013). Using cool pavements as a mitigation strategy to fight urban heat island—A review of the actual developments. *Renewable and Sustainable Energy Reviews*, *26*, 224–240.

Santamouris, M., Synnefa, A., & Karlessi, T. (2011). Using advanced cool materials in the urban built environment to mitigate heat islands and improve thermal comfort conditions. *Solar Energy*, *85*(12), 3085–3102.

Scheffer, T., Decomain, C., & Wrobel, S. (2001). Active hidden markov models for information extraction. In *International symposium on intelligent data analysis* (pp. 309–318). Springer.

Settles, B. (2009). *Active learning literature survey*: *CS technical reports*, University of Wisconsin-Madison Department of Computer Sciences.

Shuster, W. D., Bonta, J., Thurston, H., Warnemuende, E., & Smith, D. (2005). Impacts of impervious surface on watershed hydrology: A review. *Urban Water Journal*, *2*(4), 263–275.

Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., et al. (2019). High-resolution representations for labeling pixels and regions. arXiv preprint arXiv:1904.04514.

Takebayashi, H., & Moriyama, M. (2012). *Study on surface heat budget of various pavements for urban heat island mitigation*. Hindawi Publishing Corporation,

Talbot, L. A., Musiol, R. J., Witham, E. K., & Metter, E. J. (2005). Falls in young, middle-aged and older community dwelling adults: perceived cause, environmental factors and injury. *BMC Public Health*, *5*(1), 1–9.

Tao, A., Sapra, K., & Catanzaro, B. (2020). Hierarchical multi-scale attention for semantic segmentation. arXiv preprint arXiv:2005.10821.

Thomas, N. D., Gardiner, J. D., Crompton, R. H., & Lawson, R. (2020). Keep your head down: Maintaining gait stability in challenging conditions. *Human Movement Science*, *73*, Article 102676.

Thomas M. Menino, T. J. T. (2013). Boston complete streets. URL: https://tooledesign.com/project/boston-complete-streets-manual.

Tillson, G. W. (1900). *Street pavements and paving materials: a manual of city pavements: the methods and materials of their construction*. John Wiley & Sons.

Van Dam, T. J., Harvey, J., Muench, S. T., Smith, K. D., Snyder, M. B., Al-Qadi, I. L., et al. (2015). *Towards sustainable pavement systems: a reference document*: *Technical report*, United States: Federal Highway Administration.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., et al. (2020). Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wu, H., Sun, B., Li, Z., & Yu, J. (2018). Characterizing thermal behaviors of various pavement materials and their thermal impacts on ambient environment. *Journal of Cleaner Production*, *172*, 1358–1367.

Yang, J., Jin, S., Xiao, X., Jin, C., Xia, J. C., Li, X., et al. (2019). Local climate zone ventilation and urban land surface temperatures: Towards a performance-based and wind-sensitive planning proposal in megacities. *Sustainable Cities and Society*, *47*, Article 101487.

Yin, L., Cheng, Q., Wang, Z., & Shao, Z. (2015). 'Big data'for pedestrian volume: Exploring the use of google street view images for pedestrian counts. *Applied Geography*, *63*, 337–345.

Yuan, Y., Chen, X., & Wang, J. (2019). Object-contextual representations for semantic segmentation. arXiv preprint arXiv:1909.11065.

Zhang, Y., Odeh, I. O., & Han, C. (2009). Bi-temporal characterization of land surface temperature in relation to impervious surface area, NDVI and NDBI, using a sub-pixel image analysis. *International Journal of Applied Earth Observation and Geoinformation*, *11*(4), 256–264.

Zhao, S., Wang, Y., Yang, Z., & Cai, D. (2019). Region mutual information loss for semantic segmentation. arXiv preprint arXiv:1910.12037.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., & Torralba, A. (2017). Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 633–641).

Zhu, S., & Mai, X. (2019). A review of using reflective pavement materials as mitigation tactics to counter the effects of urban heat island. *Advanced Composites and Hybrid Materials*, *2*(3), 381–388.

Zhu, Y., Sapra, K., Reda, F. A., Shih, K. J., Newsam, S., Tao, A., et al. (2019). Improving semantic segmentation via video propagation and label relaxation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8856–8865).