

Measuring the Utilization of Public Open Spaces by Deep Learning: a Benchmark Study at the Detroit Riverfront

Peng Sun

Rui Hou

Jerome P. Lynch

University of Michigan

{patcivil, rayhou, jerlynch}@umich.edu

Abstract

Physical activities and social interactions are essential activities that ensure a healthy lifestyle. Public open spaces (POS), such as parks, plazas and greenways, are key environments that encourage those activities. To evaluate a POS, there is a need to study how humans use the facilities within it. However, traditional approaches to studying use of POS are manual and therefore time and labor intensive. They also may only provide qualitative insights. It is appealing to make use of surveillance cameras and to extract user-related information through computer vision. This paper proposes a proof-of-concept deep learning computer vision framework for measuring human activities quantitatively in POS and demonstrates a case study of the proposed framework using the Detroit Riverfront Conservancy (DRFC) surveillance camera network. A custom image dataset is presented to train the framework; the dataset includes 7826 fully annotated images collected from 18 cameras across the DRFC park space under various illumination conditions. Dataset analysis is also provided as well as a baseline model for one-step user localization and activity recognition. The mAP results are 77.5% for pedestrian detection and 81.6% for cyclist detection. Behavioral maps are autonomously generated by the framework to locate different POS users and the average error for behavioral localization is within 10 cm.

1. Introduction

POS are a vital part of healthy cities offering public spaces for social interactions, exercise and to enjoy with nature. The global trend of shrinking households drives a need for social contact outside the home [13]. Healthy living [1] encourages people to stay physically active within a pleasant environment (e.g. green parks). Studies [45, 12, 15] also show that physical activities which can be promoted in POS are beneficial to mental health and can substantially reduce the risk of chronic disease (e.g. cardiovascular disease, pul-

monary disease, metabolic syndromes). People-centered urban design of POS is drawing increasing attention with designed POS to promote physical activities [38] with post occupancy evaluation (POE) [42] later performed to verify design assumptions. Methods of measuring the usage of POS (e.g. counting, mapping, tracing, and test walks) are often executed manually, which are time and labor intensive [14]. Hence, an automatic sensing method is needed to investigate patrons within POS.

POS (e.g. parks and greenways) can serve as anchoring points driving the transformation of urban spaces in populated cities into more lively environments. In 2003, the Detroit Riverfront Conservancy (DRFC) was incorporated to restore the international riverfront area of Detroit (Fig. 1a). A network of 100 surveillance cameras (Fig. 1a) has been installed at the riverfront to ensure the safety of the area. These surveillance cameras provide a means of automating the assessment of patrons within POS using computer vision (CV). Recently, deep learning based-CV techniques have benefited many domains including video surveillance, assisted living, human-robot interaction, and health care. This paper presents a CV-based approach to automatically localize users and recognize their activities for measuring the usage of POS using surveillance cameras. The proposed cyber-physical-social system (Fig. 1b) can provide informative guidelines for improving urban spaces during their design. Users' activity information (e.g. usage patterns and occupancy frequencies) is extracted by CV algorithms and shared with urban designers and park managers for decision making.

The paper is organized as follows: first, a review of the existing related works on vision-based urban studies is presented; second, the detection model, behavioral mapping, and the evaluation metrics are presented; third, the collection and annotation process of a custom dataset is described; fourth, a case study of using the proposed framework to study the usage of a POS at the Detroit Riverfront is presented. The key contributions of this study are in three-fold: (1) a custom image dataset is established for user detection in POS as well as evaluation metrics, (2) a baseline model

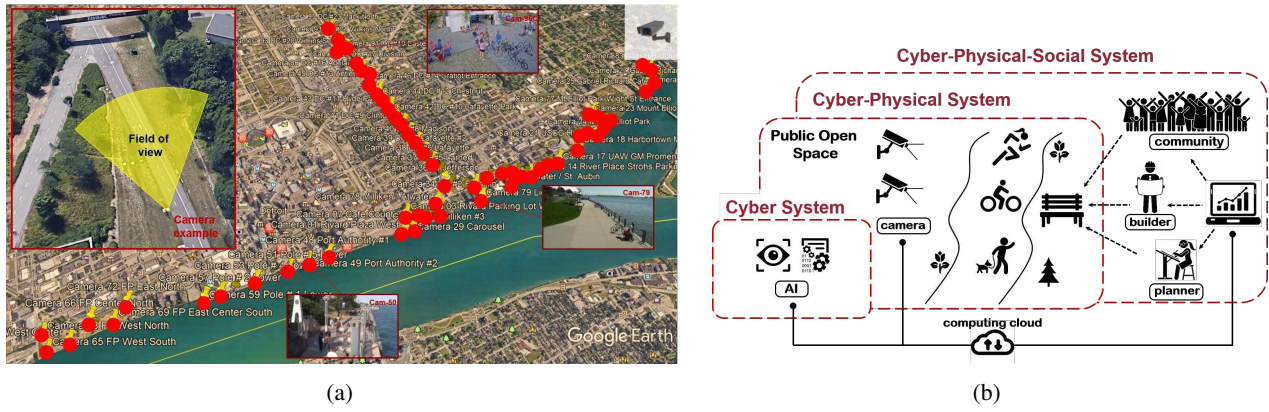


Figure 1: (a) A map of the Detroit riverfront with surveillance cameras. (b) CV-based cyber-physical-social system (CPSS) for urban planning and design.

for POS sensing is trained and evaluated for user localization and activity recognition using monocular surveillance cameras, and (3) a benchmark study is demonstrated and heat-maps of user behaviors are depicted for key park locations.

2. Related Work

Vision-based sensing in urban studies: Architects started to adopt video recordings to study people’s social behavior and interactions in public spaces since the late 1960’s [52]. Manual observation is the primary way for such a purpose. With the development of CV-based sensing technology, coupled with cheap computing power, there are emerging studies using CV to measure physical urban changes [40], to analyze pedestrian gait [22], and to study autonomous driving [2, 23] in urban streets. However, few studies have been focused on measuring the usage of POS for improving urban planning. [53] is one of the few studies that employed a computer vision-based method to measure human activity (e.g. sitting or walking) in a POS. The people detection method was based on background subtraction and blob detection. Although it introduced video sensing into urban space studies, the sensing system was unable to perform real-time detection; the detection robustness and accuracy is also presumed to suffer from adopting low-level features of images. An advanced detection method should be investigated to perform real-time, robust user sensing in POS by using deep features of images.

CV-based object detection: Object detection models are utilized to identify and locate objects in images. Region-based detection models (e.g. Fast R-CNN [16] and Faster R-CNN [44]) rely on region proposal networks (RPN) [44] and convolutional neural networks (CNN) to estimate bounding boxes (bbox) of objects. In contrast, single-stage models (e.g. YOLO [43] and SSD [35]) perform object detection without a separate region proposal step. Although the former methods suffer from comparatively slow detection speed, they outperform the latter in detection accuracy[56]. Mask R-CNN [20] is a region-based detection method that provides richer information of a de-

tected object with an additional instance segmentation besides a bbox. Furthermore, detected contours can provide location information of specific body parts [30]. Recently, new anchor-free detectors (e.g. FCOS [49], FoveaBox [26]) have been developed to achieve higher performance in detecting accurate bounding boxes without using anchor references.

Human activity recognition: For many years, researchers have been working on human activity recognition (HAR) using different kinds of sensors. Non-vision based (e.g. wearable) and vision-based sensors are two categories of sensors used. In [27, 54], multiple wearable sensors (e.g. accelerometers, gyroscopes, and magnetometers) are attached to the body of a subject to measure motion attributes in order to recognize different activities. However, wearable sensors are intrusive to users [25] and can only cover a very limited number of users in a POS. Traditional CV methods using vision-based sensors usually rely on a few visual features (e.g. HoG, local binary pattern, or RGB), which makes it difficult to achieve robust people detection (especially under extreme illumination conditions). In the past few years, deep learning [28, 19] methods using deep neural networks have grown rapidly and are drawing much attention as the result of their supreme performance in different applications. Others have used deep features of images to extract high-level representation for activity recognition tasks [4, 34, 18]. A complete HAR study [50] usually consists of at least two steps: 1) person detection, and 2) activity/body shape recognition based on feature representations (e.g. silhouette representation). Unlike past HAR studies, the objective of this study is to achieve person detection and activity recognition in a single step. To do so, the recognition of different activities will be embedded within the classification task of the instance segmentation.

Image datasets related to people detection: Image datasets have promoted research in many topics in computer vision, such as object detection, pedestrian detection, and autonomous driving. Among the datasets available for generic object detection (Table 1), ImageNet [9], COCO [33], and PASCAL VOC [11] are very popular choices

Dataset	Main purpose	# Cls	# Ppl. cls	# Annt. imgs	Inst. segm.	Cam/sensor type	Night	Rain	Locations
ImageNet-1k	object det.	1000	1 (pers.)	1.3M	No	-	No	No	-
VOC-2007	object det.	20	1 (pers.)	10K	No	-	No	No	-
COCO-2014	object det.	80	1 (pers.)	160K	Yes	-	No	No	-
CityScapes	auton. driving	19	2 (pers.+rider)	25K	Yes	veh.-mount. dash cam.	No	No	50 cities in Europe
Mapillary Vistas	auton. driving	66	5 (pers.+rider)	25K	Yes	phone, tablet, action cams, etc.	No	No	around world
nuScenes	auton. driving	23	4 (ped. etc.)	1.4M	Yes	veh.-mount.cam.+lidar+ladar	Yes	Yes	Boston, SG
Caltech Pedestrian	ped. det.	1	1 (ped.)	250k	No	veh.-mount. dash cam.	No	No	LA
Cyclist Benchmark	cyclist det.	1	1 (cycl.)	14.6K	No	veh.-mount. stereo cam.	No	No	Beijing
OPOS	user sense	15	10	7.8K	Yes	PTZ surv. cam.	Yes	Yes	Detroit

Table 1: Comparison of image datasets for object detection.

with many object classes associated with each of them. However, there are only a few “person” classes in these datasets which limit human activity classification. The earlier datasets which were designed solely for pedestrian detection (e.g. Caltech Pedestrian [10] and INRIA Person [7]) only include pedestrians. The human pose in these datasets is either standing straight or walking, while cyclists are excluded or ignored on the annotated images to avoid dealing with the similar appearance between pedestrians and cyclists. In 2016, a cyclist benchmark dataset [31] was designed for cyclist detection. However, there is only one object class (cyclist) annotated in bbox without any instance segmentations. Today, because of the emerging technologies of autonomous driving, many datasets are built from data collected from diverse set of sensors mounted on vehicles (e.g. cameras, radar, GPS, IMU, and LIDAR sensors). Because the main focus for autonomous driving is on the quick detection of pedestrians and other vehicles on the road during driving, the related datasets include only a few classes of people on the road. For example, CityScapes [6] includes pedestrian and rider, Mapillary Vistas [41] includes persons (individual and groups) and riders (motorcyclist, bicyclist, other), and nuScenes [5] includes different pedestrians (adult, child, construction worker, and police officer). Hence, in order to detect user activities that are usually observed in a POS (e.g. walking, riding a bike, riding a scooter, sitting), the aforementioned datasets would not achieve the joint task of user detection and activity recognition. A new dataset, including people with various physical activities (that often occur in POS), is needed to train a detection model for POS.

3. Methodology

3.1. Detection Model

A detection model serves as the primary building block of the sensing framework proposed herein for modeling public use of city park spaces. A user mapping algorithm

(mapping from the 2D pixel coordinate system of the camera to the 3D coordinate system of the real physical space) is built on top of the detection model to assess POS utilization. In this study, the Mask R-CNN [20] model is utilized for user detection and activity recognition. In Mask R-CNN, different CNN architectures can be adopted as the computational core to extract features from an input image. ResNet [21] is a well-known convolutional backbone to achieve object detection and segmentation. ResNet extracts multiple feature maps from an image while RPN generates regions within which objects potentially lie. The RPN slides over either one of the last convolutional feature maps or over feature maps combined during the use of a feature pyramid network (FPN) [32]. FPN serves as a feature detector on top of the feature maps generated by ResNet to gather rich information (i.e. multi-scale features) in both semantics and location. RPN will be trained together with the ResNet backbone in an end-to-end fashion.

3.2. Detection Evaluation Metrics

The accuracy of a detector can be evaluated using average precision (AP) and average recall (AR) under a certain threshold of intersection over union (IoU). IoU measures the percentage of overlap between the ground truth and the detected bounding box. In this study, AP and AR are defined following the COCO metrics [33] which are averaged over multiple IoU values from 0.5 to 0.95 with a step size of 0.05. The mean average precision (mAP) is defined as the mean value of the AP_c across different object classes, c . For a specific class, c , AP is also divided into AP_c^{sm} (small), AP_c^{med} (medium), and AP_c^{lg} (large), depending on sizes of objects within images. In order to analyze detection performance and detection error, a single AP cannot provide a thorough diagnosis for a trained detector. For instance, the effects of different types of false positives [24] may include the localization error and confusion with semantically similar objects. Hence, the popular precision-recall (PR) curve methods [3, 8] will also be utilized in this study to diagnose

possible detection errors.

3.3. Camera Calibration

The pinhole camera model [36] is used to perform camera calibration and 3D bbox estimation [39] using geometry. In this work, a scene view is formed by projecting points of a 3D location defined in world coordinates $\{X, Y, Z\}$ onto the image plane defined by the pixel coordinate system $\{u, v\}$ using a perspective transformation:

$$sm' = A [R|t] M \quad (1)$$

where A is a matrix of intrinsic parameters of the camera, $[R|t]$ is the joint rotation-translation matrix (including extrinsic parameters), and m' and M are the locations in the pixel coordinate system and world coordinate system, respectively. The intrinsic matrix A and distortion coefficients (i.e. radial and tangential distortion) can be computed by using a chessboard at various positions. The extrinsic parameters of a camera can be computed by the Levenberg–Marquardt algorithm [29] relating the locations of objects in both world and pixel coordinates (e.g. the pattern of the broken white lines on the pedestrian path).

3.4. Behavioral Mapping and Evaluation

Behavioral mapping [46] is a standard method in urban design used to evaluate how participants utilize a built environment by recording the behaviours of participants. It can provide a useful record of where people are, what they do, and how their behaviors are distributed in space. Behavioral maps consist of two forms: place-centered and individual-centered.[48] In this paper, the former type of behavioral mapping is studied and monocular surveillance cameras are used to estimate the 3D bbox and the location of users in a POS. The ground is presumed to be flat and the origin, O , is fixed to the ground plane ($Z = 0$). This restraint changes the once ill-posed problem into an optimization problem. The mapping algorithm is implemented on top of the Mask R-CNN detector allowing detection and mapping to be performed automatically on each video frame. The evaluation of the mapping can be performed by computing the average difference between real locations (precisely defined by field measurements) of reference points and the projected locations of the corresponding location in the image pixel coordinates (using calibrated camera parameters).

4. OPOS Dataset

A dataset named “Objects in Public Open Spaces” (OPOS) is prepared based on images captured using the DRFC surveillance camera network.

Scene selection: The DRFC surveillance cameras are operated at 5 or 30 fps all day (i.e. 24 hours). A total of 18 cameras (scenes) of interest are selected to col-

lect raw video data. The video resolutions vary across different cameras (e.g. 1280×720px on the Dequindre Cut, 1108×832px at Cullen Plaza). Surveillance cameras capture video frames with: various traffic types (e.g. multiple pedestrians, cyclists, scooter riders), usual and rare classes (e.g. cars/trucks, skaters, dogs), different levels of camera distortion, and various illumination and weather conditions (e.g. sunny, cloudy, and rainy as shown in Fig. 2).

Object classes: The objects of interest are classified into four (4) different super-categories (i.e. *people*, *vehicle*, *accessory* and *animal*) and 15 different object classes (as shown in Fig. 3). In order to avoid the confusion for distinguishing the correct object class for each object, rules for taxonomy are defined for annotations. For example, the definitions of the *people* classes are as follows: *pedestrian* are people observed with crucial body parts (head, shoulder, knees, and at least one foot) connected; *pedpart* are people partially observed with body parts (head, shoulder, butt and knees) connected; *cyclist* are people riding a bicycle or with a bicycle observed with parts (head, shoulder, bicycle handlebar, and at least one bicycle wheel) connected; *cycpart* are people most likely riding a bicycle partially observed with head, shoulder, crouching gesture, and bicycle handlebar/one bicycle wheel; *scooterer* are people riding a scooter; *skater* are people on a skateboard; *roller* are people riding roller skates; *sitter* are people that are sitting; *people lying* are people that are lying on the grass or ground; *peopleother* are people with other activities or with severe occlusions, but with head and shoulder observed.

Detector-aided annotation: The annotation process includes the initial collection of a seed dataset and the incremental annotation of more images by using detection models (pre-trained on the custom dataset). The seed dataset includes images on which objects of interest are first manually annotated with instance segmentations. The initial iteration of the manual annotations includes 925 images. A tight bounding box is computed by the detected contour and a label is tagged in the form of “super-category.class.obj#” (e.g. “*person.pedestrian_6*”). A detector (e.g. Mask R-CNN) fine-tuned on the seed dataset is utilized to segment objects from a new batch of images. Manual revision is fol-



Figure 2: Images from the DRFC surveillance cameras and the annotated objects under various illumination conditions: sunny (col 1), cloudy (col 2), rainy (col 3), and evening (col 4).

lowed to correct any errors. The iterations of detector-aided annotations are performed in the following steps iteratively for each 1000 additional processed images: (1) perform instance segmentation by using a newly trained Mask R-CNN detector with a threshold confidence score of 75%; (2) filter the detection (segmentation) results by a threshold area of 600 px² and import them into LabelMe (a graphical image annotation tool) JSON file [51]; (3) manually check/revise the instance segmentation (e.g. drag vertices to reshape the polygon to tightly confine a *pedestrian*) if needed; (4) store the new set of data and use it to train a new detection model.

Data format and privacy protection: Labels and annotations are stored in the COCO object detection (segmentation) format, including 4 sections (i.e. info, licenses, images, annotations). Each annotation section includes information of object id, category id, segmentation, bounding box, etc. To ensure the privacy of the dataset, the cascaded convolutional network [55] is used to detect human faces within the dataset and the results are double checked manually to cover human faces. The final results are used to blur human faces in the dataset to preserve user privacy.

Statistical analysis: The statistics of the annotations in the OPOS dataset are analyzed and presented in this section. The dataset (Table 2) includes various people, vehicles, and other mobile object classes. The total number of segmented objects is 18.9K and the most frequent objects are *pedestrian* and *cyclist*. The annotated objects are under various weather conditions (e.g. 6.7% rainy, 45.8% cloudy, and 47.4% sunny) during spring and summer seasons. Also, 14.1% of the annotations are during the evening with street lights on, and 7.2% of the annotations are small-size objects.

In real situations, surveillance cameras can capture occluded objects or truncated ones at the image boundary (due to objects moving in or out of view). Sometimes, even with occlusions or truncation, humans can still infer the object as a partially viewed object (e.g. *pedestrian* or *cyclist*) by



Figure 3: Examples of different classes within the OPOS dataset.

Object Class	# Obj.	Area (px ²)		Aspect ratio	
		mean	std	mean	std
<i>pedestrian</i>	9675	4659	4003	2.81	0.48
<i>cyclist</i>	5092	6432	6505	2.56	0.42
<i>scooter</i>	466	5056	5231	2.83	0.51
<i>skater</i>	43	5746	4043	2.99	0.53
<i>sitter</i>	704	3719	2844	1.52	0.44
<i>peopleother</i>	736	6449	6433	1.45	0.55
<i>pedpart</i>	487	8475	7063	2.25	0.54
<i>cycpart</i>	192	8321	7902	1.50	0.56
<i>roller</i>	86	4716	4119	2.55	0.43
<i>lay</i>	6	2344	1963	0.80	0.51
<i>stroller</i>	216	4955	3848	1.64	0.47
<i>car</i>	736	12861	14521	0.80	0.24
<i>vehicleother</i>	51	6000	6319	1.24	0.34
<i>dog</i>	303	1543	1070	1.60	0.57
<i>umbrella</i>	109	4449	5257	0.48	0.16

Table 2: Numbers of annotations per object class in OPOS.

referring to some visual cues (e.g. the gesture of standing up straight or the gesture of holding the steering bar of a bicycle). Hence, truncated/occluded people classes (e.g. denoted as *pedpart* and *cycpart*) are also included in the dataset to check whether a detection model can distinguish between visual cues.

5. Experiments

5.1. Dataset Setup for Training

The OPOS custom dataset includes 7826 images. Pre-training on public datasets is preferred before performing transfer learning on the OPOS dataset. The weights of the ResNet backbones are first pre-trained on the ImageNet-1k dataset. The weights of most of the models (except the first model in Table 4) are additionally pre-trained on the coco.2017_train dataset (pretrained weights are from the Detectron website [17]). The training and test sets of the OPOS dataset are split by the ratio of 9:1.

Fine-tuning schedule: the pre-trained weights are fine-tuned on the OPOS dataset using the maskrcnn-benchmark platform [37] with a NVIDIA 1070 GPU. The mini batch size is set as 2 images/batch, and horizontal flipping is adopted for data augmentation. The schedule including 90K iterations starts at a learning rate of 0.0025, decreases by a factor of 0.1 after both 60K and 80K iterations, and finally terminates at 90k iterations. This schedule results in 25.56 epochs of training on the training data set. The ResNet backbone has five stages [21] (denoted as conv1-5) with each stage consisting of several stacked residual blocks. The first two stages (i.e. conv1 and conv2) have been frozen, while the weights of the other stages are kept flexible for updating.

5.2. Study of Detection Models

Treatment of similar people classes: There are 10 classes within the *people* super-category. However, there are many similarities across closely related classes within

Treatment of Occl./Trun. Classes	bbox /segm	mAP _c per ppl. class								mAP (ppl.)	mAP (overall)
		<i>ped.</i>	<i>cycl.</i>	<i>scoot.</i>	<i>skater</i>	<i>sitter</i>	<i>other</i>	<i>pedpart</i>	<i>cycpart</i>		
Merging (part to whole)	<i>bbox</i>	77.5%	81.6%	74.7%	63.3%	65.7%	53.4%	NA	NA	69.4%	67.7%
	<i>segm</i>	74.7%	81.2%	74.3%	66.4%	59.4%	53.9%	NA	NA	68.3%	66.2%
Filtering (part to pplother)	<i>bbox</i>	78.5%	81.6%	81.4%	57.4%	64.7%	61.5%	NA	NA	70.9%	68.4%
	<i>segm</i>	75.4%	81.8%	79.7%	65.1%	59.4%	61.3%	NA	NA	70.5%	67.0%
Separating (keep part classes)	<i>bbox</i>	78.3%	81.4%	77.4%	53.6%	64.9%	56.1%	58.4%	38.0%	63.5%	64.6%
	<i>segm</i>	75.2%	81.6%	76.6%	50.5%	59.3%	55.5%	59.0%	37.6%	61.9%	62.7%

Table 3: Comparison of models with the same backbone (ResNet50-FPN) trained on the same datasets (Imagenet+COCO+OPOS) with different treatments of occluded/truncated classes.

the same super-category; for example, a *roller* (people wearing roller skates) is very similar to a *pedestrian* except slight differences on the foot wear. If they are treated as two separate classes for training, the detection results are expected to be inconsistent. For example, a *roller* would be detected incorrectly as a *pedestrian* (most likely to happen) in far field, but once the person is approaching the camera, the person would be identified as *roller* once the model spotted the nuances in foot wear. This would cause trouble in practice for user detection or future tracking tasks. Hence, in the current stage of research, *roller* is merged into *pedestrian*.

Treatments of occluded/truncated classes: The occluded/truncated classes (i.e. *pedpart* and *cycpart*, denoted as part classes) consist of 3.6% of the overall objects. There is a need to study the influence of different treatments of the part classes. Three arrangements are made here: (1) “merging” – treating the part classes as the corresponding whole classes (i.e. *pedpart* is treated as *pedestrian*, and *cycpart* is treated as *cyclist*); (2) “filtering” – treat the part classes as *peopleother* class (i.e. both *pedpart* and *cycpart* are treated as *peopleother*); (3) “separating” – treating the part classes as individual classes (i.e. *pedpart* and *cycpart* classes are kept as separate classes). As shown in Table 3, the detection performances are very close for merging and filtering treatments, and the two outperform the separating treatment by more than 3% in mAP. This might be due to the strong similarity between the truncated classes and the corresponding whole classes. The results show that the treatment of either filtering or merging reduces the confusion in detection for these classes, and thus improves the detection performance. Although the merging treatment slightly underperforms the filtering treatment (e.g. 1.3% lower in bbox mAP and 0.8% lower in segmentation mAP), the merging treatment is more logical in understanding and practical usage. Hence, the rest of the study will be performed using the merging treatment of the part classes.

Datasets	bbox /segm	mAP _c per ppl. class						mAP (ppl.)
		<i>ped.</i>	<i>cycl.</i>	<i>scoot.</i>	<i>skat.</i>	<i>sitter</i>	<i>other</i>	
imagenet+bbox		76.1%	80.3%	76.3%	48.5%	65.3%	49.3%	66.0%
opos	segm	73.3%	80.4%	78.3%	44.3%	56.4%	49.4%	63.7%
imagenet+bbox		77.5%	81.6%	74.7%	63.3%	65.7%	53.4%	69.4%
coco+opos	segm	74.7%	81.2%	74.3%	66.4%	59.4%	53.9%	68.3%

Table 4: Comparison of models with the same backbone (ResNet50-FPN) trained on different datasets.

Effect of pre-training: Some of the models are directly fine-tuned on the OPOS dataset by using the pre-trained weights on the ImageNet dataset (denoted as ImageNet+OPOS) and the other models use pre-trained weights on both ImageNet and coco.2017_train datasets (denoted as ImageNet+COCO+OPOS). As shown in Table 4, the APs for various people (ppl.) classes observe a general increase with the additional pre-training on the coco.2017_train dataset. The increases for mAP of the overall people classes are 3.4% in bbox detection and 4.6% in segmentation detection, respectively. Hence, pre-training on more datasets would benefit the ultimate performance of the detection models.

Impact of backbone architectures: The detection models of Mask R-CNN with different backbone architectures are trained on ImageNet+COCO+OPOS, as shown in Table 5. The results show that ResNet50-FPN (“FPN” denotes a head from the second half of FPN) outperforms ResNet50-C4 (“C4” denotes a head from the fourth module of ResNet50) with an increase of 2.9% in bbox detection and 4.6% in segmentation detection for overall classes. The results show the advantages of using combined feature maps at different scales (from FPN head) over using a single feature map (from C4 head). Backbone architectures with various depths (e.g. ResNet50-FPN and ResNet101-FPN) are also compared (where “50” and “101” refer to the numbers of convolutional layers). ResNet50-FPN outperforms ResNet101-FPN with the same training protocol (90k iterations). A prolonged training schedule (180k iterations) improves ResNet101-FPN by 4.3% in bbox detection.

5.3. Error Diagnosis of the Detection Model

The evaluation of a trained baseline model of Mask R-CNN is demonstrated here. The detection task requires

Backbone architecture	bbox /segm	mAP ^{size} for overall cls.			
		small	medium	large	all-size
ResNet50-C4	<i>bbox</i>	50.4%	66.4%	70.9%	64.8%
	<i>segm</i>	48.0%	62.0%	77.3%	61.6%
ResNet50-FPN	<i>bbox</i>	57.6%	67.8%	77.4%	67.7%
	<i>segm</i>	50.2%	66.3%	81.6%	66.2%
ResNet101-FPN	<i>bbox</i>	53.8%	66.0%	73.8%	63.8%
	<i>segm</i>	49.4%	63.6%	74.0%	61.8%
ResNet101-FPN (2x train)	<i>bbox</i>	53.3%	70.0%	76.0%	68.1%
	<i>segm</i>	49.8%	66.8%	77.2%	65.5%

Table 5: Comparison of Mask R-CNN models with different backbones (trained on ImageNet+COCO+OPOS).

detecting 11 object classes (6 classes of people with part classes merged into whole classes) which are a subset of all 15 classes in the OPOS dataset. The trained detector is evaluated on the test dataset consisting of 783 images. If not stated specifically, AP is referred as bbox AP in the rest of the study. The mAP metrics for each people class, overall people classes, and overall classes are shown in Table 3. PR curves are used to study the errors for the detection model in detail, as presented in Fig. 4 [24]. The area under each PR curve represents the AP under a specific criteria. “C75” represents the PR curve at IoU=0.75; “C50” represents the PR curve at IoU=0.50; “Loc” represents the PR curve at IoU=0.10, when localization error is ignored (denoted as “perfect localization”), but not duplicate detection; “Sim” represents the PR (IoU=0.10 for “Sim”, “Oth”, “BG”, and “FN”) after super-category false positives (class confusion within same super-category) are removed; “Oth” represents the PR after all intra-class false positives (both within super-categories and cross super-categories, denoted as “class confusion”) are removed; “BG” represents the PR after all background false positives are additionally removed; “FN” represents the PR after false negatives are additionally removed. Among all the PR plots, results of the overall object classes and *people* classes are of the most interest to evaluate the proposed detection model.

Detection performance for overall objects: The overall $AP^{0.75}$ of the trained Mask R-CNN detector is 81.3% and perfect localization would improve AP to 90.9%. While the effects of class confusion (2.5% within super-categories) and background confusion (0.6%) are very trivial compared to the effect of perfect localization. In general, the APs for small objects (e.g. 63.1% in $AP^{0.75}$) are poorer than those for both medium (e.g. 81.8% in $AP^{0.75}$) and large (e.g. 89.2% in $AP^{0.75}$) objects. For small object detection (Fig. 4c), there is a considerable amount of localization errors (13.1%) and false negatives (20.4%). Future improvement of the detection model could focus on reducing localization error and false negatives. For example, more small object annotations could be added to OPOS. However, detecting small objects intrinsically is a hard problem, because small objects include much less information (pixel-wise) and sometimes cannot provide necessary distinctive features for class detection. For medium and large object detection, the overall performances are satisfactory with plum shaped PR curves and high APs. The potential improvement could be targeted at reducing localization (LOC) error for medium objects (currently 10.7% LOC error) and reducing the false negative (FN) predictions for large objects (currently 9.5% FN error).

Detection performance for people: No substantial drawback is observed for people detection in general. The detection performance for some common people classes, for example, newly merged *pedestrian* class (including *roller*

and *pedpart*), newly merged *cyclist* class (including *cyc-part*), and *peopleother* class can be seen in Fig. 4d, 4f, and 4h, respectively. The precision for predicting *pedestrian* and *cyclist* classes are high with $AP^{0.75}$ of 92.6% and 95.4%, respectively. For *peopleother* (as shown in Fig. 4h), $AP^{0.75}$ is relatively low with a value of 59.2%. The top three effects on *peopleother* come from localization errors (19.6%), class confusion within the *people* super-category (5.6%) and false negatives (12.9%). The results stem from the definition rule for the *peopleother* class. The definitions of the first five *people* classes (i.e. *pedestrian*, *cyclist*, *scooter*, *skater*, *sitter*) are very clear with distinctive features while *peopleother* is classified as a severely truncated or occluded class that cannot be recognized as any one of the five *people* classes. Hence, the lack of explicitly defined features causes trouble in object localization and distinguishing the correct class.

5.4. Behavioral Mapping Application

The locations of different users along the Dequindre Cut (a pedestrian path in the riverfront park) are projected to road maps (width=4.5m, length=32m) as shown in Fig. 5.

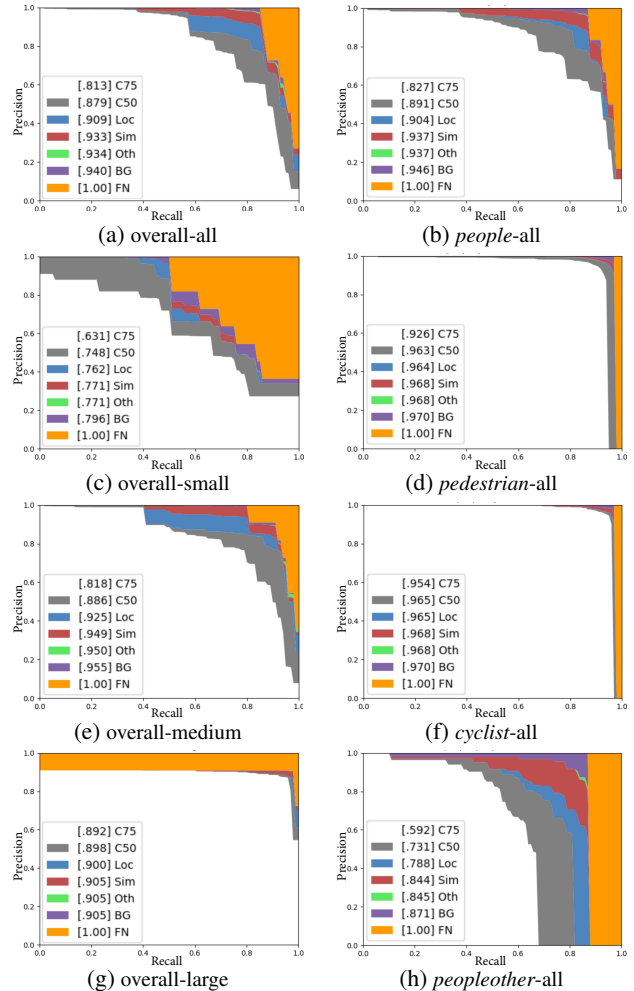


Figure 4: Precision-recall curves of (left col.) overall classes at different sizes, and (right col.) specific people classes at all sizes.

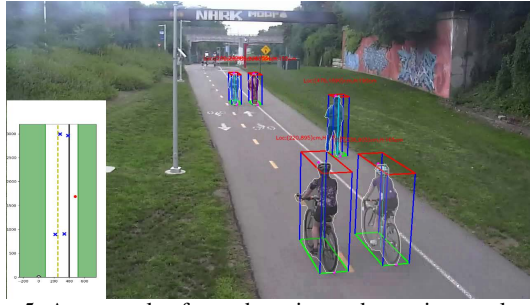


Figure 5: An example of user detection and mapping results on the Dequindre Cut. *pedestrian* is noted as ●, *cyclist* is denoted as ×, top and bottom pixels are denoted by ·.

The location and height of the 3D bounding boxes are estimated by using detected locations of object parts (e.g. feet, heads, bicycle tire) and *a priori* defined horizontal sizes of pedestrians ($w=50\text{cm}$, $l=60\text{cm}$) and cyclists ($w=50\text{cm}$, $l=160\text{cm}$). It is found that 3D bboxes of moving *pedestrian* and *cyclist* can be estimated by using monocular cameras with some assumptions (e.g. flat ground and presumed geometry).

A camera at Cullen Plaza (Fig. 6a) is calibrated by using a checkerboard and 19 reference points on the ground (shown in Fig. 6b). The error of the mapping is obtained by computing the difference between field measurements of the reference points and the projection of the corresponding image pixels. The averaged error for a static scene is 7.68cm, which is accurate enough for urban design studies in a large POS. User localization (Fig. 7a) and behavioral mappings (Fig. 7b) at any specific moment can be obtained in 0.16 s/frame. However, a statistical study over a period (e.g. density map) can provide more insights to understand the utilization of POS.

Long-term monitoring of users (*people* detections are filtered) is achieved by generating density maps using accumulated detection data over space and time. The density maps over one day (e.g. 2019-06-25, 9:00am to 17:00, in Fig. 7c) and one week (e.g. from 2019-06-24 to 2019-06-30, 9:00 to 17:00, in Fig. 7d) are generated by using kernel density estimation [47] on the detection results (frames are sampled at 1 fps). It is found that users tend to stay closer to the edges (e.g. fountain steps, poles, fence area, statues, etc.) in a POS. For example, *sitter* tend to sit on fountain steps and *pedestrian* tend to stay near fence poles. The scenario is consistent with the “edge effect” that has been observed by architects and urban planners in the past [13].

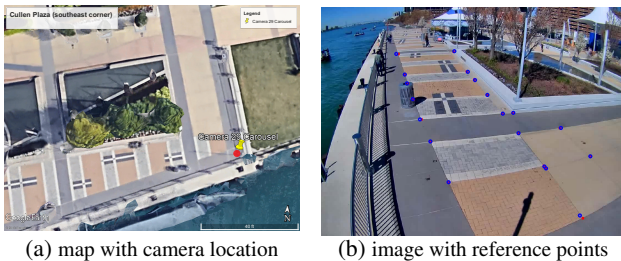


Figure 6: Map of Cullen Plaza and camera calibration with reference points. ●: selected pixels on image, ○: projection of the world locations onto pixel coordinates using camera parameters.

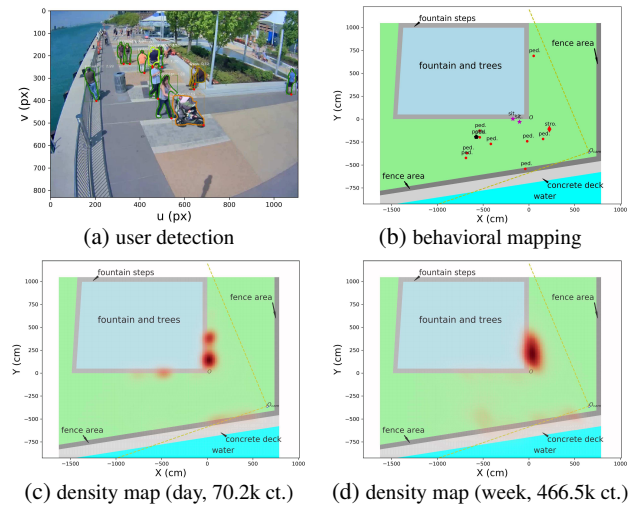


Figure 7: Examples of (a) user detection, (b) behavioral mapping results, (c) daily density map, and (d) weekly density map on the southeast corner of Cullen Plaza.

6. Conclusion

In this paper, the OPOS dataset is presented along with a baseline detection model of Mask R-CNN. The custom dataset is specifically designed for user detection in POS using surveillance cameras. A benchmark study of user sensing and behavioral mapping at the Detroit Riverfront Conservancy (DRFC) is demonstrated, providing general guidelines in building a CV-based framework to measure usage of POS for many urban design applications. The detection results show that the baseline detector has a bbox mAP of 68.4% (67.0% in segmentation) for overall objects, and a bbox mAP of 70.9% (70.5% in segmentation) for the *people* super-category. The bbox $AP^{0.75}$ for the most two common people classes (*pedestrian* and *cyclist*) that appear at the Detroit Riverfront are 92.6% and 95.4%, respectively. The averaged error for the behavioral mapping task is 7.68 cm which is satisfactory for urban studies in large public spaces. To spur future research on CV-based measurement of POS usage, evaluation metrics and an error diagnosis method is also provided to analyze the detection models. In the end, a case study of the proposed method is performed to measure the utilization of a popular plaza at the Detroit Riverfront in a week. In the future, the study would serve as a stepping stone to other challenging tasks (e.g. user counting, tracking, re-id tasks) that are associated with urban planning studies.

Acknowledgement

The support from the National Science Foundation (NSF) under grant #1831347 is gratefully acknowledged.

References

- [1] R. Arena, S. Bond, R. O'Neill, D. R. Laddu, A. P. Hills, C. J. Lavie, and A. McNeil. Public park spaces as a platform to promote healthy living: introducing a healthpark concept. *Progress in Cardiovascular Diseases*, 60(1):152–158, 2017.
- [2] D. Barnes, W. Maddern, and I. Posner. Find your own way: Weakly-supervised segmentation of path proposals for urban autonomy. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 203–210. IEEE, 2017.
- [3] M. Buckland and F. Gey. The relationship between recall and precision. *Journal of the American Society for Information Science*, 45(1):12–19, 1994.
- [4] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970, 2015.
- [5] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019.
- [6] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.
- [7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In C. Schmid, S. Soatto, and C. Tomasi, editors, *International Conference on Computer vision & Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE Computer Society, 2005.
- [8] J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine learning*, pages 233–240. ACM, 2006.
- [9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.
- [10] P. Dollár, C. Wojek, P. Perona, and B. Schiele. Pedestrian detection: A new benchmark. pages 304–311, 2009.
- [11] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.
- [12] J. Francis, L. J. Wood, M. Knuiman, and B. Giles-Corti. Quality or quantity? exploring the relationship between public open space attributes and mental health in perth, western australia. *Social Science & Medicine*, 74(10):1570–1577, 2012.
- [13] J. Gehl. *Cities for people*. Island press, 2013.
- [14] J. Gehl and B. Svarre. *How to study public life*. Island press, 2013.
- [15] B. Giles-Corti, J. F. Sallis, T. Sugiyama, L. D. Frank, M. Lowe, and N. Owen. Translating active living research into policy and practice: one important pathway to chronic disease prevention. *Journal of Public Health Policy*, 36(2):231–243, 2015.
- [16] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.
- [17] R. Girshick, I. Radosavovic, G. Gkioxari, P. Dollár, and K. He. Detectron. <https://github.com/facebookresearch/detectron>, 2018.
- [18] G. Gkioxari, R. Girshick, P. Dollár, and K. He. Detecting and recognizing human-object interactions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8359–8367, 2018.
- [19] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.
- [20] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017.
- [21] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.
- [22] H. Hediyeh, T. Sayed, M. H. Zaki, and G. Mori. Pedestrian gait analysis using automated computer vision techniques. *Transportmetrica A: Transport Science*, 10(3):214–232, 2014.
- [23] S. Hoermann, M. Bach, and K. Dietmayer. Dynamic occupancy grid prediction for urban autonomous driving: A deep learning approach with fully automatic labeling. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2056–2063. IEEE, 2018.
- [24] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. In *European Conference on Computer Vision*, pages 340–353. Springer, 2012.
- [25] A. Jalal, Y.-H. Kim, Y.-J. Kim, S. Kamal, and D. Kim. Robust human activity recognition from depth video using spatiotemporal multi-fused features. *Pattern Recognition*, 61:295–308, 2017.
- [26] T. Kong, F. Sun, H. Liu, Y. Jiang, and J. Shi. Foveabox: Beyond anchor-based object detector. *arXiv preprint arXiv:1904.03797*, 2019.
- [27] O. D. Lara and M. A. Labrador. A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys & Tutorials*, 15(3):1192–1209, 2012.
- [28] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
- [29] K. Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, 2(2):164–168, 1944.
- [30] J. Li, J. Zhao, Y. Wei, C. Lang, Y. Li, T. Sim, S. Yan, and J. Feng. Multiple-human parsing in the wild. *arXiv preprint arXiv:1705.07206*, 2017.
- [31] X. Li, F. Flohr, Y. Yang, H. Xiong, M. Braun, S. Pan, K. Li, and D. M. Gavrila. A new benchmark for vision-based cyclist detection. In *2016 IEEE Intelligent Vehicles Symposium (IV)*, pages 1028–1033. IEEE, 2016.
- [32] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection.

- In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2117–2125, 2017.
- [33] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.
 - [34] J. Liu, Y. Li, S. Song, J. Xing, C. Lan, and W. Zeng. Multi-modality multi-task recurrent neural network for online action detection. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 2667–2682, 2018.
 - [35] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37. Springer, 2016.
 - [36] Y. Ma, S. Soatto, J. Kosecka, and S. S. Sastry. *An invitation to 3-d vision: from images to geometric models*, volume 26. Springer Science & Business Media, 2012.
 - [37] F. Massa and R. Girshick. maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch. <https://github.com/facebookresearch/maskrcnn-benchmark>, 2018.
 - [38] G. R. McCormack, M. Rock, A. M. Toohey, and D. Hignell. Characteristics of urban parks associated with park use and physical activity: A review of qualitative research. *Health & Place*, 16(4):712–726, 2010.
 - [39] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka. 3d bounding box estimation using deep learning and geometry. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7074–7082, 2017.
 - [40] N. Naik, S. D. Kominers, R. Raskar, E. L. Glaeser, and C. A. Hidalgo. Computer vision uncovers predictors of physical urban change. *Proceedings of the National Academy of Sciences*, 114(29):7571–7576, 2017.
 - [41] G. Neuhold, T. Ollmann, S. Rota Buló, and P. Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 4990–4999, 2017.
 - [42] W. F. Preiser, E. White, and H. Rabinowitz. *Post-Occupancy Evaluation (Routledge Revivals)*. Routledge, 2015.
 - [43] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.
 - [44] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015.
 - [45] J. F. Sallis, M. F. Floyd, D. A. Rodríguez, and B. E. Saelens. Role of built environments in physical activity, obesity, and cardiovascular disease. *Circulation*, 125(5):729–737, 2012.
 - [46] H. Sanoff and G. Coates. Behavioral mapping: An ecological analysis of activities in a residential setting. *International Journal of Environmental Studies*, 2(1-4):227–235, 1971.
 - [47] B. W. Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.
 - [48] R. Sommer and R. Sommer. *A practical guide to behavioral research: Tools and techniques*, volume 5. Oxford University Press New York, 1997.
 - [49] Z. Tian, C. Shen, H. Chen, and T. He. Fcos: Fully convolutional one-stage object detection. *arXiv preprint arXiv:1904.01355*, 2019.
 - [50] M. Vrigkas, C. Nikou, and I. A. Kakadiaris. A review of human activity recognition methods. *Frontiers in Robotics and AI*, 2:28, 2015.
 - [51] K. Wada. labelme: Image Polygonal Annotation with Python. <https://github.com/wkentaro/labelme>, 2016.
 - [52] W. H. Whyte. *The social life of small urban spaces*. Washington, D.C.: Conservation Foundation, 1980.
 - [53] W. Yan and D. A. Forsyth. Learning the behavior of users in a public space through video tracking. In *2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION’05)-Volume 1*, volume 1, pages 370–377. IEEE, 2005.
 - [54] A. Y. Yang, S. Iyengar, S. Sastry, R. Bajcsy, P. Kuryloski, and R. Jafari. Distributed segmentation and classification of human actions using a wearable motion sensor network. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. IEEE, 2008.
 - [55] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.
 - [56] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu. Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 2019.