BEV-Net: Assessing Social Distancing Compliance by Joint People Localization and Geometric Reasoning

Zhirui Dai[1] Yuepeng Jiang[1] Yi Li[1] Bo Liu[2] Antoni B. Chan[3] Nuno Vasconcelos[1]

[1]Department of Electrical and Computer Engineering, UC San Diego
[2]Wormpex AI Research
[3]Department of Computer Science, City University of Hong Kong
{zhdai,yuj009,yil898,boliu}@eng.ucsd.edu, abchan@cityu.edu.hk, nvasconcelos@ucsd.edu

Abstract

Social distancing, an essential public health measure to limit the spread of contagious diseases, has gained significant attention since the outbreak of the COVID-19 pandemic. In this work, the problem of visual social distancing compliance assessment in busy public areas, with wide field-of-view cameras, is considered. A dataset of crowd scenes with people annotations under a bird's eye view (BEV) and ground truth for metric distances is introduced, and several measures for the evaluation of social distance detection systems are proposed. A multi-branch network, BEV-Net, is proposed to localize individuals in world coordinates and identify high-risk regions where social distancing is violated. BEV-Net combines detection of head and feet locations, camera pose estimation, a differentiable homography module to map image into BEV coordinates, and geometric reasoning to produce a BEV map of the people locations in the scene. Experiments on complex crowded scenes demonstrate the power of the approach and show superior performance over baselines derived from methods in the literature. Applications of interest for public health decision makers are finally discussed. Datasets, code and pretrained models are publicly available at GitHub[1].

1. Introduction

Social distancing, the strategy of maintaining a safe distance between people in public spaces, has been shown to be an effective measure against the transmission of contagious pathogens, including influenza virus and coronavirus [57, 7, 23]. However, the monitoring of social distancing by human observers is neither practical in many settings nor scalable. This has motivated an interest in methods to detect and count social distancing violations automatically. While non-vision-based methods are available [2], they typically require users to install certain applications on their mobile devices, and are limited in precision of distance estimates.

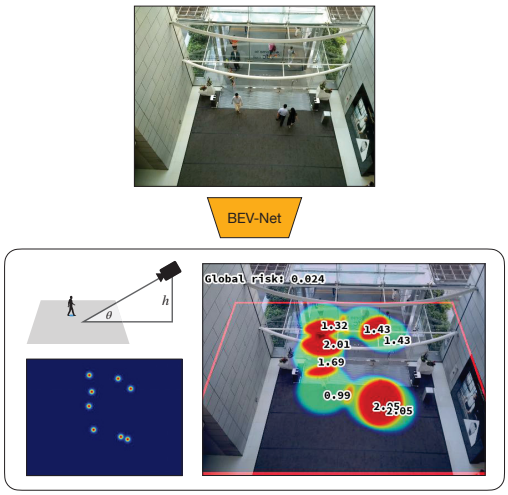

Figure 1: Set of tasks proposed for social distancing compliance assessment. Given an input image, models are expected to jointly predict camera geometry (**top left**), bird's eye view heatmap of person locations (**bottom left**), local risk map highlighting areas with high probability of infection (**right**), as well as individual and global risk level for the entire scene (annotated over the risk map).

Computer vision offers a viable alternative for the collection of social distance measurements. In particular, it has several advantages for the monitoring of public spaces. First, it can leverage surveillance cameras that are already available in many public locations. No expensive infrastructure changes are required.

Second, anonymization of visual data is straightforward by removing all facial identities, as the system has no access to other sensitive information of pedestrians. This makes it much more privacy preserving than the monitoring of mobile devices, or similar approaches. On the other hand,

[1]https://github.com/daizhirui/BEVNet

[2]The *DP-3T* contact tracing protocol [59], for example, estimates distances using Bluetooth signal on smartphones.

it can produce complete statistics of social distancing violations, as there is no prerequisite on the population (e.g. smartphone with Bluetooth enabled). While not suited for contact-tracing, these statistics can be very useful to decision makers, e.g. to enable the implementation of highly localized "lock-downs," time-varying control of pedestrian access to certain areas, etc.

Computer vision has a long history of sensing humans in images. Object detection [14, 17, 16, 52, 37] and instance segmentation [8, 21, 35] recognize and localize objects by bounding boxes or pixel-wise masks. While effective for close objects and sparsely populated scenes [12, 39], detection quality degrades substantially for far cameras and busy spaces, involving significant degrees of occlusion. In fact, on such scenes, it is even impossible to collect accurate bounding box annotations. Social distancing measuring is more closely related to crowd-counting methods [66, 5, 42], which are trained to produce a heatmap that highlights the head location of every person in the scene. However, because these methods don't explicitly reason about scene geometry, they are unsuitable to estimate distances between individuals. Furthermore, head locations are not ideal for estimating scene distances, since heads do not lie on a shared plane in 3D, due to people height variation. Much more accurate distance estimates can usually be obtained by reconstructing the scene ground plane and measuring feet distances. This, however, presents additional challenges, due to occlusion.

In this work, we consider the problem of social distancing compliance assessment (SDCA), which aims to measure the distances between individuals in a scene and detect violations of social distancing thresholds. Based on the above observations, we argue that SDCA requires a geometry-aware approach, where the relevant extrinsic parameters of the camera are estimated and used to generate a bird's eye view (BEV) map of people locations, as illustrated in Figure 1. We propose a novel benchmark for SDCA, CityUHK-X-BEV, which repurposes the CityUHK-X dataset [28] to the SDCA problem, by adding ground-plane annotations. Specifically, for each head location in the dataset, the corresponding feet position is annotated, and mapped to the BEV, using the known intrinsic and extrinsic camera parameters. A novel set of evaluation criteria is introduced, each focusing on a different aspect of the task: *Localization* metric that measures the accuracy in detecting real-world locations of people in the ground plane; *local risk* metric that evaluates the capability to discover regions with high chance of infection; *global risk* metric that predicts the overall risk level of captured scene. Figure 1 shows examples of these tasks. Unlike many vision problems that localize object in the image plane, these geometry-aware criteria directly evaluate the capacity of models to make predictions in the 3D ground plane, leading to outputs that are much more informative for real-world applications that require metric information.

A multi-branch convolutional architecture, BEV-Net, is then proposed to solve the SDCA task. BEV-Net follows an encode-decoder structure, using a projective transformation module to convert convolutional feature maps from image view to BEV. The decoder is implemented with a pose regression branch that estimates camera parameters and three separate branches to predict feet, head and BEV heatmaps. These branches are trained with individual losses, in a multi-task manner. To compensate for the height variations in the crowd, a group transformation module with spatial self-attention is used to group people by head height and independently align the feet and head feature maps of the resulting groups. Experiments show that the BEV-Net outperforms all DET and CC baselines under all proposed SDCA evaluation metrics. It is further shown, through ablation experiments, that both head and feet annotations are essential to achieve the best prediction quality.

A number of applications of potential interest for public health decision makers are then illustrated. These range from the characterization of risks for a single image, as illustrated in Figure 1, to global measures of scene risk, integrated over image datsets, as shown in Figures 7 and 8. The latter can be used to identify events of unusually large risk or inform the deployment of risk mitigation measures, such as the introduction of obstacles in the scene to modify walking patterns and other crowd behaviors.

The paper makes four major contributions: First, we introduce the idea of using computer vision for joint geometric reasoning and social distancing compliance assessment on public spaces. Second, a novel benchmark for SDCA in crowd scenes, CityUHK-X-BEV, is introduced with person-level annotations in bird's eye view. Third, a multi-branch convolutional network, BEV-Net, is proposed and shown to achieve best SDCA results by learning to perform both heatmap prediction and geometric reasoning. Finally, we show promising results for several potential applications of the SDCA framework in the public health domain.

2. Related Work

Object detection (DET). Object detection methods recognize and localize multiple classes of objects with bounding boxes. While early algorithms relied on hand-crafted visual features [44, 9], the introduction of CNN-based detectors [17, 16, 52, 4, 51, 38] trained on large-scale image databases [10, 39] has enabled dramatic performance gains.

Existing approaches to SDCA have mostly relied on pre-trained DET models, making few technical advances to their architectures. [13] proposed to detect social distancing violations by regressing head and feet locations from the bounding box of each detected person. While effective for sparsely populated environments, such an approach does not scale to busy spaces and distant cameras with a large field of view, as is usually the case of large public spaces.

In addition, [13] requires external homography calculation based on markers manually placed in the scene. By contrast, the proposed BEV-Net automatically estimates camera geometry and is able to estimate social distances for crowded environments and wide field of view cameras.

Crowd counting (CC). Crowd counting focuses on counting the number of people in an image. State-of-the-art methods either learn to regress head counts from the image data directly [46], or predict a people density map which is then integrated to obtain the people count [66, 56, 5, 36, 40]. Crowd scene datasets tend to be collected in public spaces and focus on busy scenes [63, 25, 60, 28], where the estimation of head locations is difficult due to small people sizes and significant occlusion. These are the scenes that we emphasize in this work, where we augment a popular crowd counting dataset with rich annotations required for SDCA.

Most CC methods are trained to produce density maps in the camera view, a simpler problem than the proposed combination of SDCA tasks. An exception is WACC [64], which learns to predict ground-plane heatmap directly. The BEV-Net differs from this work in three main aspects: First, WACC requires inputs from multiple cameras, while BEV-Net is designed to work with a single camera view. Second, WACC is supervised by head annotations only, leading to inaccurate ground-plane locations due to varying head heights; BEV-Net addresses this by producing feet annotations that, unlike heads, lie on the common ground plane. Third, WACC assumes known geometry for all camera views, while BEV-Net jointly learns to predict extrinsic camera parameters.

Geometry in computer vision. The scene geometry recovery is a classical problem in computer vision [20]. This can be decomposed into the estimation of camera parameters and scene geometry, i.e., depth variability in different parts of the scene. Since the introduction of deep learning, both components have been estimated by neural networks, typically by using two dedicated branches that are trained jointly, in an end-to-end manner [32, 67, 18, 27]. The scene reconstruction required by SDCA amounts to recovering the ground plane and 3D feet locations of all individuals. This is not trivial because feet locations are frequently occluded in the camera view. BEV-Net addresses this by leveraging head locations and the regularizing geometric constraint that a standing person's head and feet are co-located in BEV.

Social behavior analysis. Computer vision methods have been applied to modeling human behavior in public spaces. One line of work achieves this through the task of trajectory prediction [34, 49, 29, 1, 19], which requires generating plausible motion paths for pedestrians in the image plane. Early work used physics models such as Social Force [22] to account for interaction between humans [47, 50, 62, 53]; more recently, neural network modules were used to capture such dependencies between agents, e.g. with recurrent networks [1, 3, 55] or graph convolutional networks [48, 58]. Other tasks for social behavior modeling have also been explored, including early action detection [54, 30, 45] and group activity recognition [6, 33, 24]. All of the above tasks require temporal modeling on video data and semantic understanding of human activities; this work instead focuses on sensing the spatial locations of people, which can be efficiently recovered from individual image frames.

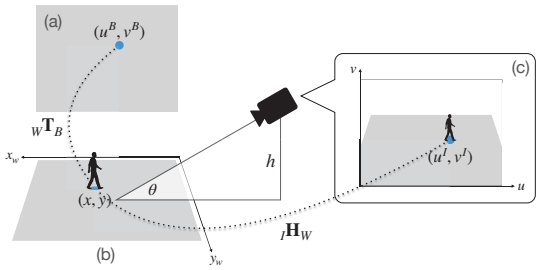

Figure 2: Homography between BEV, IV, and world coordinates. **(a)** BEV map with person location (u^B, v^B); **(b)** world model with camera at height h, pitch angle θ and person at position (x, y) on the ground plane; **(c)** image captured by camera with feet detected at pixel location (u^I, v^I). Coordinates are converted across views through homography matrices ${}_W\mathbf{T}_B$ and ${}_I\mathbf{H}_W$; gray area denotes region of interest.

3. Social Distancing Compliance Assessment

To the best of our knowledge, no prior work has attempted to evaluate the quality of visual SDCA in busy public spaces and large field-of-view scenes. In this section, we propose a new dataset for this task.

3.1. Motivation

A dataset for SDCA should satisfy various requirements. First, it should contain a wide range of scenes with varying people densities. Second, it should include ground-truth locations for the people in the scene, either in 3D world coordinates or in the form of a BEV heatmap. Optionally, it could provide ground-truth for the intrinsic and extrinsic camera parameters, allowing direct camera pose supervision during training and facilitating the recovery of ground plane and homography between camera view and BEV.

DET datasets [11, 39, 65] are not suitable for SDCA since the number of people per image tends to be low. CC datasets are more relevant, as they contain abundant examples of people gathering in clusters—often within 1 to 2 meters, the range of droplet transmission [61, 15, 31]—and the scene/camera configurations are most suitable for monitoring social distances in public spaces. However, geometric meta-information is unavailable in most CC datasets.

CityUHK-X [28] is an exception, providing extrinsic camera parameters, including height and pitch angle, which make it a potential SDCA benchmark. Nevertheless, it has limitations. Like other CC datasets, it only provides image annotations for each *head* in the scene. Even with known camera parameters, image head locations aren't enough to

Figure 3: Example ground-truth heatmaps overlaid on the input image. **Left**: IV maps, with heads in green and feet in red. Region of interest is bounded by a red perspective box. **Right**: BEV location map. Centers and bottom centers of heatmaps are aligned.

recover locations in world coordinates, as each individual's height is variable and unknown. Hence, additional annotations are needed for the *feet* locations of each person in the image, as well as head-feet correspondences.

3.2. Annotation

The Amazon MTurk platform was used to collect feet annotations for CityUHK-X with correspondences between feet and heads. Given a crop of the scene image with annotated head location, MTurk workers were asked to annotate the center point between both feet, and to verify if the feet were clearly visible (feet location is precise) or occluded (feet location is estimated). 87,746 feet locations were annotated in the 2,982 scene images in CityUHK-X. Among them, 63,669 (72.6%) were clearly visible and 24,077 (27.4%) occluded. Detailed description of the annotation procedure can be found in the supplemental.

3.3. Ground-truth Generation

As illustrated in Figure 3, ground-truth is provided in the form of three density maps: feet and head maps in image view (IV maps), and person location map in BEV coordinates (BEV map).

Image view maps. Following [66], the ground-truth head and feet locations in image view are represented with heatmaps $\mathbf{M}_{\text{head}}$ and $\mathbf{M}_{\text{feet}}$ composed by mixture of Gaussians. Each Gaussian from the head (feet) heatmap is centered at the annotated head (feet) location of a person in the image, with a fixed standard deviation $\sigma = 5$ pixels.

Bird's eye view map. Given the image coordinates of annotated keypoints and camera geometry, the BEV map $\mathbf{M}_{\text{BEV}}$ is generated using a homography. This is achieved by projecting the keypoints to world coordinates, choosing a suitable region of interest within the ground plane, then resampling into a fixed-size BEV heatmap, as illustrated in Figure 2. First, the world coordinate frame is defined by setting the ground plane to $z = 0$ and the camera sensor to $(0, 0, h)$, where h denotes camera height above ground. The camera's yaw angle is zero in this setting. Further assuming that the camera has pitch angle θ, and zero roll (the roll angle could be zero by transforming the input image properly), the transformation between world coordinates (x, y) in the ground plane and image coordinates (u^I, v^I) is then given by the homography (derivations in supplemental)

$$\begin{bmatrix} u^I & v^I & 1 \end{bmatrix}^\top = {}_I\mathbf{H}_W \begin{bmatrix} x & y & 1 \end{bmatrix}^\top, \tag{1}$$

where ${}_I\mathbf{H}_W$ is constructed as a function of h, θ and intrinsic camera parameters such as focal lengths (f_u, f_v).

Second, to define a proper region of interest (RoI) on the ground plane, we require that the center (u_c^B, v_c^B) of BEV map be projected to the image center (u_c^I, v_c^I), and the bottom-center pixel in BEV be aligned with the bottom-center image pixel, as indicated in 3. The scale factor s, measuring distance in meters spanned per BEV pixel, is then given by

$$s = 2(x_c - x_{bc})/H = h/\left[\beta(v_c^I\alpha + f_v\beta)\right] \tag{2}$$

where H is the height of the BEV map. The transformation from BEV map coordinates (u^B, v^B) to world coordinates in the ground plane is then given by the homography

$$\begin{bmatrix} x & y & 1 \end{bmatrix}^\top = {}_W\mathbf{T}_B \begin{bmatrix} u^B & v^B & 1 \end{bmatrix}^\top, \tag{3}$$

where ${}_W\mathbf{T}_B$ is constructed to align the BEV map with image center, and rescale to s meters per BEV pixel. Finally, the homography between image and BEV map coordinates is obtained by combining (1) and (3) into

$$\begin{bmatrix} u^I & v^I & 1 \end{bmatrix}^\top = {}_I\mathbf{H}_B \begin{bmatrix} u^B & v^B & 1 \end{bmatrix}^\top, \tag{4}$$

where

$${}_I\mathbf{H}_B = {}_I\mathbf{H}_W \, {}_W\mathbf{T}_B. \tag{5}$$

Given a set of feet locations $\{\mathbf{q}_j\}_{j=1}^d$ in image $\boldsymbol{I}$, the corresponding locations in BEV coordinates are given by $\{{}_I\mathbf{H}_B{}^{-1}\mathbf{q}_j\}_{j=1}^d$. Similar to the IV maps, the BEV map $\mathbf{M}_{\text{BEV}}$ is generated using a Gaussian kernel with $\sigma = 5$ px. Example heatmaps are shown in figure 3.

3.4. Evaluation protocol

We propose two types of criteria for evaluating the quality of predicted BEV heatmaps for SDCA.

Localization error. Models are required to identify BEV locations in real-world distance units (e.g. meters or feet), for all individuals in the scene. Given a set of predicted locations $\hat{\mathbb{X}} = \{\hat{\mathbf{x}}_i\}_{i=1}^M$ and a set of ground-truth locations $\mathbb{X} = \{\mathbf{x}_i\}_{i=1}^N$, the Chamfer distance [2]

$$D(\hat{\mathbb{X}}, \mathbb{X}) = \frac{1}{M}\sum_{i=1}^{M} \min_j \|\hat{\mathbf{x}}_i - \mathbf{x}_j\| + \frac{1}{N}\sum_{j=1}^{N} \min_i \|\hat{\mathbf{x}}_i - \mathbf{x}_j\| \tag{6}$$

is used to evaluate localization error in terms of real-world distances. Predicted locations $\hat{\mathbb{X}}$ are determined from the BEV heatmap, using non-maximum suppression of size 5×5 pixels, followed by pixel-wise thresholding at the heatmap value 10^{-3}. The non-zero entries of the post-processed BEV map are then extracted and converted to world coordinates using (3). We also evaluate the normalized chamfer distance $D_n(\hat{\mathbb{X}}, \mathbb{X}) = \frac{D(\hat{\mathbb{X}}, \mathbb{X})}{2d_0}$, which measures the localization error as a percent of the safe distance

threshold. A minimum safe distance of $d_0 = 1.5$m is used, but can be adjusted per public health guidelines.

Risk estimation error. Models are expected to measure compliance with social distancing by estimating risk levels in the scene, either locally or globally. *Local* risk levels are represented as a heatmap $\mathbf{R}$ on the ground plane, with greater values indicating locations with higher risk of infection. Risk is estimated from the BEV localization map $\mathbf{M}_{\text{BEV}}$ of sect. 3.3 by applying a scale-adaptive kernel $\mathbf{K}$ determined by a chosen infection risk model.

$$\mathbf{R}(u, v) = \mathbf{M}_{\text{BEV}}(u, v) * \mathbf{K}(u, v) . \tag{7}$$

When the infection risk is defined as simply the number of people within the safe distance to a person, $\mathbf{K}$ is a disk-shaped kernel of radius $r = d_0/s$, where s is the scale parameter of equation 2. Therefore, $\mathbf{R}(u, v)$ represents the count of people within radius r of location (u, v). Evaluating the accuracy of the risk map by comparing pixel-wise values can be sensitive to overcrowded areas and fail to capture borderline cases. Instead, we pose local risk estimation as a segmentation problem: For a given image, the network outputs a binary mask by thresholding the risk heatmap, with positive regions indicating areas where transmission is likely to occur; the prediction quality is evaluated by intersection-over-union (IoU) w.r.t. ground-truth mask.

Global risk levels are defined by counting occurrences of social distancing violation—the total number of people that fail to maintain a minimum distance d_0 from one another—and normalizing by the area covered by the BEV map. This is estimated by multiplying the BEV heatmap with the binary risk mask, then integrating over the space

$$R_g(r_0) = \frac{1}{s^2 HW} \sum_{\substack{u,v \\ R(u,v) \geq r_0}} M_{\text{BEV}}(u, v) \tag{8}$$

where r_0 is a threshold on the acceptable risk level. Risk above r_0 is considered unsafe, indicating the possibilty of infection due to violation of social distancing recommendations. Global risk error is measured by mean squared error (MSE) between estimated and ground-truth risks.

Figure 1 shows sample outputs for the tasks described above. The BEV heatmap $\mathbf{M}_{\text{BEV}}$ relies on accurate detection of each individual, while local and global risk estimates are more robust to minor localization errors, as the influence of each person is spread out in the ground plane.This diverse set of criteria assures that model outputs perform well with respect to both metrics of interest for computer vision (localization) and public health practitioners (risk levels).

4. BEV-Net

In this section we present BEV-Net, a unified framework for the solution of crowd counting, camera pose estimation and social distancing compliance assessment.

4.1. Multi-branch Encoder-Decoder

The design of BEV-Net is based on the encoder-decoder architecture commonly used in CC models [5, 42]. However, we have found that directly training an encoder-decoder to generate BEV heatmaps leads to poor results, since a fully convolutional architecture has difficulty modelling the large and non-uniform displacements that exist between a pixel location in the input image and the corresponding location in the BEV map.

BEV-Net addresses this problem through the multi-branch architecture of figure 4. The head and feet branches are trained to predict heatmaps for head and feet in the image view (IV), respectively. Standard convolutional encoder and decoder layers suffice to implement these branches, as the input image and the IV heatmaps are aligned. For SDCA, these heatmaps are not of interest per se, since they contain no metric information. However, the addition of the two branches enables supervision for head and feat locations, which is critical to let the network select which image features to pay attention to. In this sense, they can be seen as a top-down attention mechanism.

All metric information is recovered by the central BEV branch. This branch has two stages. The first is a pose regression network that runs in parallel with head and feet branches, enabling geometric reasoning by learning to predict the height h and pitch angle θ of the camera. This is implemented with a CNN feature extractor, followed by layers of MLP, and supervised by a pose-estimation loss. The second stage uses the camera parameters to rectify the IV head and feet feature maps, denoted $\mathbf{F}_{\text{IV, head}}$ and $\mathbf{F}_{\text{IV, feet}}$, into BEV coordinates. Given the predicted camera pose $(\hat{h}, \hat{\theta})$, the IV feature maps are first aligned in BEV space through projective transformation T_{head} and T_{feet} (details in section 4.2 and 4.3). This produces a pair of feature maps $\mathbf{F}_{\text{BEV, head}}$ and $\mathbf{F}_{\text{BEV, feet}}$ in BEV, which are then concatenated along channel dimension and fed to the BEV decoder, eventually producing the predicted BEV map.

4.2. BEV-Transform: Feature-level Homography

The projective transformation between IV and BEV (see figure 3) creates a spatially varying displacement between IV and BEV feature map locations. This makes it difficult to predict the BEV map from the IV feature maps, since the convolution operation is not naturally suited to model spatially varying displacements. Inspired by [26], we address this problem by designing a differentiable BEV Homography transformation module based on (4), called BEV-Transform, to perform feature-level homography mapping.

Given the predicted camera pose $(\hat{h}, \hat{\theta})$, for a plane at height h_0, BEV-Transform calculates the homography transformation ${}_I\hat{\mathbf{H}}_B = H(\hat{h} - h_0, \hat{\theta})$ that maps BEV map grid $\mathbf{G}_B$ into the IV map sample grid $\mathbf{G}_I$:

$$\begin{bmatrix} u^I & v^I & 1 \end{bmatrix}^\top = {}_I\hat{\mathbf{H}}_B \begin{bmatrix} u^B & v^B & 1 \end{bmatrix}^\top , \tag{9}$$

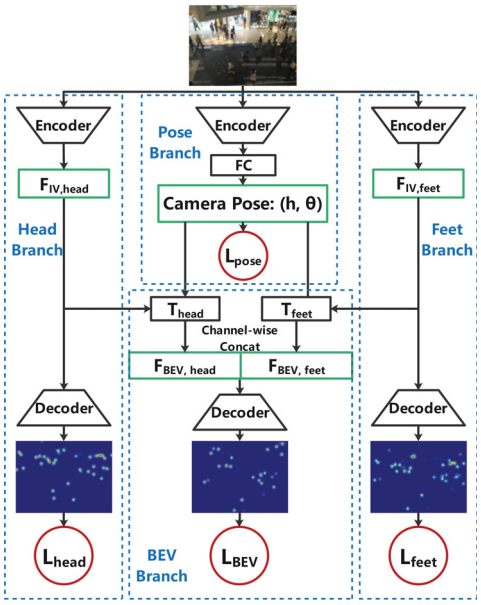

Figure 4: The BEV-Net architecture is a multi-branch model that jointly performs geometric reasoning (**pose** branch), image-view keypoint localization (**head** & **feet** branch), and bird's eye view person localization (**BEV** branch).

where (u^B, v^B) and (u^I, v^I) are coordinates in $\mathbf{G}_B$ and $\mathbf{G}_I$ respectively. Note that h_0 is different for head and feet planes. Hence, two matrices $({}_I\hat{\mathbf{H}}_B)_i, i \in \{\text{head, feet}\}$ are needed to transform the feature maps of head and feet, respectively, in order to align them in BEV. In practice, more matrices are used as revealed in Section 4.3.

Given these matrices, (9) is then used to transform the feature maps from image to BEV coordinates, using

$$(\mathbf{F}_{\text{BEV},i})_{u^B,v^B} = (\mathbf{F}_{\text{IV},i})_{u^I,v^I}, \quad i \in \{\text{head, feet}\} \tag{10}$$

As in [26], this is implemented with a differentiable bilinear interpolation layer. It should be noted that when feature maps are internally resized by the network, the coordinates (u_c^I, v_c^I) of image center and focal lengths (f_u, f_v) are scaled proportionally to match the map size.

4.3. Group BEV-Transform with Spatial Attention

The determination of feet and head plane heights has different complexity. For feet, the height is determined trivially as $h_0 = 0$. For heads, however, since people's heights are different, the head annotations are not in the same horizontal plane in the world frame. One possibility would be to simply ignore head locations. However, the co-location of the vertical projections of head and feet is a strong regularization constraint for camera pose estimation.

To take advantage of this, BEV-Net relies on a set of head planes that quantify the range of person heights. To cover both adult and child heights, planes are placed at heights 1.1m, 1.2m, ..., 1.8m from the ground. People in different regions of the image are then automatically assigned to different height planes by a self-attention mechanism, as illustrated in figure 5. The IV head feature maps $(\mathbf{F}_{\text{IV}})_{head}$

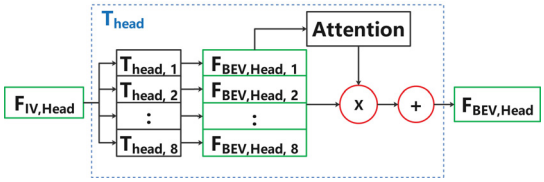

Figure 5: Architecture of Grouping BEV transform modules and spatial attention mechanism

are first mapped by the homographies $({}_I\hat{\mathbf{H}}_B)_{head,i}, i \in \{1, \ldots, 8\}$, associated with the different head heights, into a set of BEV head feature maps $\{(\mathbf{F}_{\text{BEV}})_{head,i}\}$. An attention branch consist of three convolutional layers is then used to compute a weight map $\mathbf{W}_i$ per BEV head feature map. The set of weight maps $\{\mathbf{W}_i\}$ are then normalized by a 2D softmax layer, notated as $\{\tilde{\mathbf{W}}_i\}$. The resulting spatially varying weighted combination of the feature maps $\sum_{i=1}^{8} (\mathbf{F}_{\text{BEV}})_{head,i} * \tilde{\mathbf{W}}_i$ is finally used as the BEV head feature map $(\mathbf{F}_{\text{BEV}})_{head}$.

4.4. Multi-task Loss

As shown in figure 4, BEV-Net is trained with four loss functions. The loss functions for the head, feet and BEV branches are all MSE losses:

$$L_i = \text{MSE}(\hat{\mathbf{M}}_i, m\mathbf{M}_i) + \alpha \text{MSE}(g(\frac{\hat{\mathbf{M}}_i}{m}), g(\mathbf{M}_i)) \tag{11}$$

where $\alpha = 1.25 \times 10^{-6}$, $\quad i \in \{\text{head, feet, BEV}\}$, $g(\mathbf{A}) = \sum_{u,v} A_{u,v}$ gives the global people counting. We amplify ground-truth heatmaps by a factor of $m = 100$ for better convergence, following practices in [66]. The pose loss combines two MSE losses, for camera height and pitch angle:

$$L_{\text{pose}} = \lambda_{\text{angle}}(\hat{\theta} - \theta)^2 + \lambda_{\text{height}}(\hat{h} - h)^2 \tag{12}$$

where λ_{angle} and λ_{height} are the weight factors. The final loss is a weighted sum of the above four loss functions,

$$L = \lambda_{\text{BEV}} L_{\text{BEV}} + \lambda_{\text{head}} L_{\text{head}} + \lambda_{\text{feet}} L_{\text{feet}} + L_{\text{pose}} \tag{13}$$

The loss weights are set to $\lambda_{\text{height}} = 0.02, \lambda_{\text{angle}} = 2.0$, $\lambda_{\text{head}} = \lambda_{\text{feet}} = 1.0$, and $\lambda_{\text{BEV}} = 8.0$ in all experiments.

5. Experiments

In this section we present experimental evaluations of SDCA on the CityUHK-X-BEV dataset.

5.1. Experimental setup

Training procedure. The head, feet and pose branches are pretrained for 50 epochs before training the BEV-Net model. All the training process uses *AdamW* [43] with learning rate $lr = 0.0008$ exponentially decreasing by factor 0.98 per epoch. A batch size of 8 and a train-validation split of 4:1 was used in all experiments. After pre-training, the BEV-Net is trained end-to-end. The BEV branch is first trained for 5 epochs with frozen pre-trained branches. All branches are then unfrozen step by step and jointly trained for 195 epochs.

Type	Architecture	Pretrain	BEV Heatmap *MSE* $\times 10^{-7}\downarrow$	Localization *Chamfer* $D \times 1\text{m}$ / $D_n\%\downarrow$	Local risk map *IoU%* $\uparrow$	Global risk *MSE* $\times 10^{-4}\downarrow$
DET	Mask R-CNN [21] †	COCO [39]	3.89	2.26 / 75.3	46.63	43.35
	Faster R-CNN [52]		2.71	1.53 / 51.0	67.01	9.91
	CSP [41] †	CityPersons [65]	4.21	4.57 / 152.3	28.75	62.26
CC	CSRNet [36]	NWPU-Crowd [60]	4.03	5.49 / 183.0	26.80	57.72
	DSSINet [40]	ShanghaiTech B [66]	4.94	3.95 / 131.7	29.71	51.01
	IV-Net (Head)	CityUHK-X [28]	5.36	3.90 / 130.0	30.01	40.19
	IV-Net (Feet)		8.25	3.65 / 121.7	22.41	9.03
	CC oracle	—	5.61	3.51 / 117.0	30.17	48.13
SDCA	BEV-Net (Ours)	CityUHK-X-BEV	**1.34**	1.25 / 41.7	**71.25**	**6.24**
	- Feet only		1.38	1.26 / 42.0	68.12	7.62
	- Head only		2.03	1.32 / 44.0	67.43	8.95
	- No group transf.		1.36	**1.24 / 41.3**	69.65	7.08

Table 1: SDCA performance, tested on CityUHK-X-BEV. Under BEV-Net: "Head/feet only" removes feet/head branch from model (fig. 4); "No group transf." uses a plain head branch w/o group transforms (sect. 4.3). *CC oracle* uses ground-truth head heatmap and camera pose. † denotes model evaluated without fine-tuning.

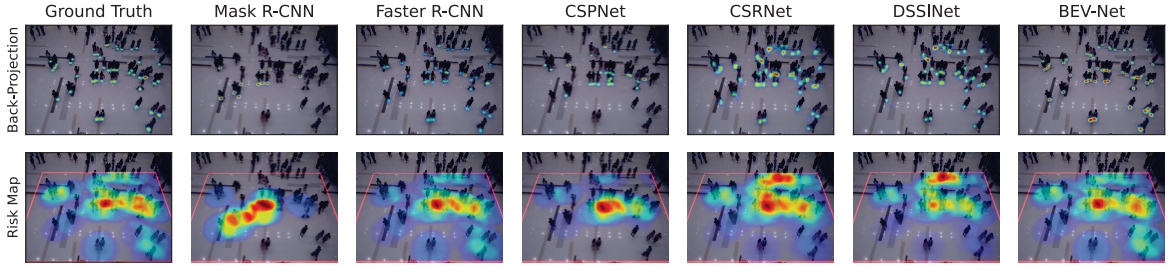

Figure 6: BEV localization & risk heatmaps from different models. Both maps are back-projected to image space for visualization.

Comparisons. We compare BEV-Net to two types of baselines. A *Detection*-based approach (DET) utilizes a person detector combined with the pretrained pose branch used by BEV-Net. The bottom center of each bounding box is used as the feet location $\mathbf{q}_i^j$. These locations are then converted to world coordinates in BEV using ${}_I\boldsymbol{H}_B{}^{-1}\mathbf{q}_i^j$ with projection matrix ${}_I\boldsymbol{H}_B$ estimated from the predicted camera pose (see section 3.3). We use pretrained CSPNet [41], Mask R-CNN [21] and Faster R-CNN [52]. The Faster R-CNN is finetuned on pseudo ground-truth bounding boxes generated from head/feet locations with aspect ratio $1/(3\cos\theta)$.

A *Counting*-based approach (CC) uses standard crowd-counting networks to generate IV head heatmaps, which we project to BEV using the BEV-Transform and the same pose branch used in DET methods. We use four networks: finetuned CSRNet [41] and DSSINet [40]; and our IV-Net, which is the head/feet branch in BEV-Net. The vertical displacement between head locations and the ground plane is compensated for by subtracting an average pedestrian height used in [28]—1.75 meters—from the predicted camera height. To explore the upper bound of counting-based methods for SDCA, we introduce a *CC oracle* which uses the ground-truth camera pose parameters and head maps.

Various ablations of the proposed BEV-Net are also evaluated. To study the effect of feet and head branches, we remove each of them from the architecture, leading to "Head only" and "Feet only" variants. We further ablate the head branch by replacing its group BEV-Transform module with a naive BEV homography ("No group transf.").

5.2. Quantitative results

Model performance. Table 1 summarizes the performance of all methods in terms of the evaluation metrics of Section 3.4. The BEV-Net outperforms all detection- and crowd counting-based methods by a significant margin. Among different evaluation criteria, local and global risk estimates,and normalized Chamfer distance showed the most significant difference between models: BEV-Net obtains over 25% higher IoU in local risk prediction, $7\times$ lower error in global risk, and a 20% reduction in normalized Chamfer Distance. While detection methods like Faster R-CNN can achieve relatively good localization performance with tight bounding boxes after finetuning, their risk estimates remain unreliable due to low recall from missed pedestrians. CC networks like DSSINet suffer from poor ground-plane localization, which hurts their SDCA performance in all criteria. Notably, even *CC oracle* with ground-truth head locations and camera pose or the IV-Net that predicts feet heatmap fails to meet the accuracy of the BEV-Net due to poor localization performance or feet occlusion. This performance gap indicates that ground-plane modeling is essential for SDCA tasks, which cannot be effectively addressed by conventional detection or crowd counting approaches that operate solely in camera view.

Ablation study. Also reported in Table 1 are ablated variants of BEV-Net. First, both "Head only" and "Feet only" models performed worse than the multi-branch BEV-Net.

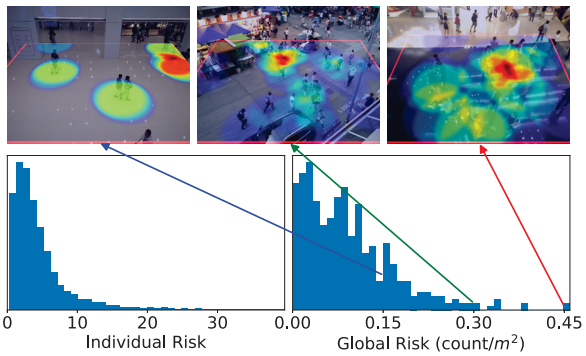

Figure 7: **Top**: Example scenes of different global risk level. **Bottom**: Histogram of individual and scene risks in test set.

This suggests that both *feet* and *head* locations are important for SDCA, which is intuitive: Head locations do not lie on a shared 2D plane due to height variations among the crowd, making it challenging to estimate real-world distances; feet locations lie on the ground plane, but are often occluded in the camera view. Among the two, feet modeling is most effective. The "Head only" model struggles to produce accurate BEV maps and localization results. This is likely to be the case for all but very crowded scenes, with large amounts of feet occlusion. Second, the group BEV-Transform module of section 4.3 enabled considerable gain over the vanilla implementation ("No group transf.") under IoU of local risk prediction and global risk error. This confirms that modeling height variations within the population is beneficial for transforming image coordinates into BEV with high precision.

5.3. Qualitative results

We next evaluate the visual quality of the BEV-Net results, and discuss its possible applications.

BEV heatmaps. Figure 6 compares BEV heatmaps and risk maps predicted by different methods. Mask R-CNN [21], Faster R-CNN [52] and CSPNet [41] suffer from low recall of people detection, such that the risk maps fail to identify all regions with high risk of infection. CSRNet [36] and DSSINet [40] can capture more risky areas, but are unable to predict the risk level correctly due to the ambiguity in head heights. BEV-Net produces the closest localization and risk heatmaps to ground-truth. Note how it accurately predicts the risk "hot-spots" inside clusters of people.

Risk-based retrieval. The multi-modal outputs from BEV-Net enable retrieval of images, individuals and clusters with highest risk of infection. Figure 7 shows the distribution of individual and scene risks in the dataset and examples of events of different risk level. Individual risk is measured by the local risk level at the ground-plane location of each person. The graph shows that only 30% of detected individuals are in compliance with the social distancing rule which prevents *two or more people* from gathering together (*i.e.* risk ≥ 2). Under a less restrictive rule that relaxes the threshold

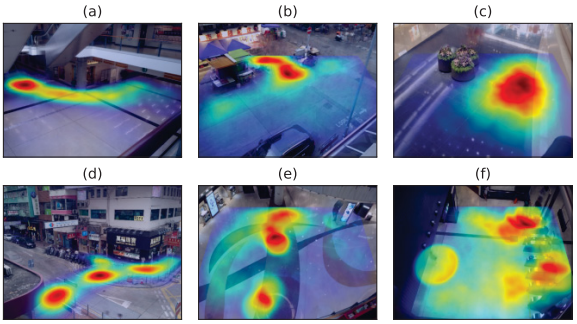

Figure 8: Mean risk map for test scenes, showing that people tend to gather at escalator entrances, near pedestrian crossings, vendor stalls, etc. Some results merit further investigation, e.g. why people prefer to use the gates near the two ends in (f).

to *five people*, the compliance rises to 72%. We believe this type of analytics is of interest for public health experts, e.g to estimate transmission factors in real world scenes. Similarly, the global risk measure can be used to detect events of high risk, where viral transmission is most likely to occur.

Scene risk analysis. While we have so far focused on image measurements, BEV-Net can also be used to estimate intrinsic risk profiles of scenes. Figure 8 shows the average risk map, over the test set, of several scenes. It can be observed that high-risk areas coincide with entrances, corners, passages or escalators. These risk maps could be used by public health decision makers to identify potential infection hot-spots and place obstacles or warning signs in the scene to mitigate infection risks.

6. Conclusion

In this work, we have introduced the problem of social distancing compliance assessment on busy public spaces, from wide field-of-view cameras, without the need for manual camera calibration or introduction of scene markers. A novel benchmark was proposed for this problem, where models are evaluated on their capability to localize people in the ground plane through geometric reasoning, and to identify regions where social distancing is violated. A multi-branch architecture, BEV-Net, was then presented, which fuses information from head and feet annotations to generate a BEV reconstruction of pedestrian locations. Experiments have shown that BEV-Net exceeds baseline methods under all evaluation metrics. Several applications of interest for public health decision makers have also been discussed.

Acknowledgements. This work was partially funded by NSF awards IIS-1924937, IIS-2041009, NVIDIA GPU donations, and a gift from Amazon. We also acknowledge and thank the use of the Nautilus platform for some of the experiments discussed above. ABC acknowledges support from a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 11212518).

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016. 3

[2] Harry G Barrow, Jay M Tenenbaum, Robert C Bolles, and Helen C Wolf. Parametric correspondence and chamfer matching: Two new techniques for image matching. In *IJCAI*, 1977. 4

[3] Federico Bartoli, Giuseppe Lisanti, Lamberto Ballan, and Alberto Del Bimbo. Context-aware trajectory prediction. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 1941–1946. IEEE, 2018. 3

[4] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 2

[5] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. Scale aggregation network for accurate and efficient crowd counting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750, 2018. 2, 3, 5

[6] Wongun Choi, Khuram Shahid, and Silvio Savarese. Learning context for collective activity recognition. In *CVPR 2011*, pages 3273–3280. IEEE, 2011. 3

[7] Charles Courtemanche, Joseph Garuccio, Anh Le, Joshua Pinkston, and Aaron Yelowitz. Strong social distancing measures in the united states reduced the covid-19 growth rate: Study evaluates the impact of social distancing measures on the growth rate of confirmed covid-19 cases across the united states. *Health Affairs*, pages 10–1377, 2020. 1

[8] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3150–3158, 2016. 2

[9] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005. 2

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[11] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):743–761, 2011. 3

[12] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 2

[13] Matteo Fabbri, Fabio Lanzi, Riccardo Gasparini, Simone Calderara, Lorenzo Baraldi, and Rita Cucchiara. Inter-homines: Distance-based risk estimation for human safety, 2020. 2, 3

[14] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2009. 2

[15] Aaron Fernstrom and Michael Goldblatt. Aerobiology and its role in the transmission of infectious diseases. *Journal of pathogens*, 2013, 2013. 3

[16] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 2

[17] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 2

[18] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency, 2017. 3

[19] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018. 3

[20] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, USA, 2 edition, 2003. 3

[21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2, 7, 8

[22] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995. 3

[23] Solomon Hsiang, Daniel Allen, Sébastien Annan-Phan, Kendon Bell, Ian Bolliger, Trinetta Chong, Hannah Druckenmiller, Luna Yue Huang, Andrew Hultgren, Emma Krasovich, et al. The effect of large-scale anti-contagion policies on the covid-19 pandemic. *Nature*, 584(7820):262–267, 2020. 1

[24] Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1971–1980, 2016. 3

[25] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2547–2554, 2013. 3

[26] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks, 2016. 5, 6

[27] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 3

[28] Di Kang, Debarun Dhar, and Antoni Chan. Incorporating side information by adaptive convolution. In *Advances in*

Neural Information Processing Systems, pages 3867–3877, 2017. 2, 3, 7

[29] Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert. Activity forecasting. In *European Conference on Computer Vision*, pages 201–214. Springer, 2012. 3

[30] Hema S Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):14–29, 2015. 3

[31] Jasmin S Kutter, Monique I Spronken, Pieter L Fraaij, Ron AM Fouchier, and Sander Herfst. Transmission routes of respiratory viruses among humans. *Current opinion in virology*, 28:142–151, 2018. 3

[32] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks, 2016. 3

[33] Tian Lan, Leonid Sigal, and Greg Mori. Social roles in hierarchical models for human activity recognition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1354–1361. IEEE, 2012. 3

[34] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer graphics forum*, volume 26, pages 655–664. Wiley Online Library, 2007. 3

[35] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2359–2367, 2017. 2

[36] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1091–1100, 2018. 3, 7, 8

[37] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2

[38] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2

[39] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 3, 7

[40] Lingbo Liu, Zhilin Qiu, Guanbin Li, Shufan Liu, Wanli Ouyang, and Liang Lin. Crowd counting with deep structured scale integration network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1774–1783, 2019. 3, 7, 8

[41] Wei Liu, Shengcai Liao, Weiqiang Ren, Weidong Hu, and Yinan Yu. High-level semantic feature detection: A new perspective for pedestrian detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 7, 8

[42] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5099–5108, 2019. 2, 5

[43] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 6

[44] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 2

[45] Shugao Ma, Leonid Sigal, and Stan Sclaroff. Learning activity progression in lstms for activity detection and early detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1942–1950, 2016. 3

[46] Mark Marsden, Kevin McGuinness, Suzanne Little, and Noel E. O'Connor. Fully convolutional crowd counting on highly congested scenes, 2017. 3

[47] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 935–942. IEEE, 2009. 3

[48] Abduallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14424–14432, 2020. 3

[49] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th International Conference on Computer Vision*, pages 261–268. IEEE, 2009. 3

[50] Stefano Pellegrini, Andreas Ess, and Luc Van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *European conference on computer vision*, pages 452–465. Springer, 2010. 3

[51] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[52] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 2, 7, 8

[53] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European conference on computer vision*, pages 549–565. Springer, 2016. 3

[54] Michael S Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *2011 International Conference on Computer Vision*, pages 1036–1043. IEEE, 2011. 3

[55] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1349–1358, 2019. 3

[56] Deepak Babu Sam, Shiv Surya, and R Venkatesh Babu. Switching convolutional neural network for crowd counting. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4031–4039. IEEE, 2017. 3

[57] Jane D Siegel, Emily Rhinehart, Marguerite Jackson, and Linda Chiarello. Guideline for isolation precautions: preventing transmission of infectious agents in healthcare settings 2007, 2007. 1

[58] Jianhua Sun, Qinhong Jiang, and Cewu Lu. Recursive social behavior graph for trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 660–669, 2020. 3

[59] Carmela Troncoso, Mathias Payer, Jean-Pierre Hubaux, Marcel Salathé, James Larus, Edouard Bugnion, Wouter Lueks, Theresa Stadler, Apostolos Pyrgelis, Daniele Antonioli, et al. Decentralized privacy-preserving proximity tracing. *arXiv preprint arXiv:2005.12273*, 2020. 1

[60] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting. *arXiv preprint arXiv:2001.03360*, 2020. 3, 7

[61] William F Wells et al. On air-borne infection. study ii. droplets and droplet nuclei. *American Journal of Hygiene*, 20:611–18, 1934. 3

[62] Kota Yamaguchi, Alexander C Berg, Luis E Ortiz, and Tamara L Berg. Who are you with and where are you going? In *CVPR 2011*, pages 1345–1352. IEEE, 2011. 3

[63] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. Cross-scene crowd counting via deep convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 833–841, 2015. 3

[64] Qi Zhang and Antoni B Chan. Wide-area crowd counting via ground-plane density maps and multi-view fusion cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8297–8306, 2019. 3

[65] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3221, 2017. 3, 7

[66] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 589–597, 2016. 2, 3, 4, 6, 7

[67] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 3