International Journal of Computer Vision manuscript No.
(will be inserted by the editor)

Synthesizing a Scene-Specific Pedestrian Detector and Pose Estimator for Static Video Surveillance

Can we learn pedestrian detectors and pose estimators without real data?

Hironori Hattori* · **Namhoon Lee*** ·
Vishnu Naresh Boddeti · **Fares Beainy** · **Kris M. Kitani** · **Takeo Kanade**

Received: date / Accepted: date

Abstract We consider scenarios where we have zero instances of real pedestrian data (*e.g.*, a newly installed surveillance system in a novel location in which no labeled real data or unsupervised real data exists yet) and a pedestrian detector must be developed prior to any observations of pedestrians. Given a single image and auxiliary scene information in the form of camera parameters and geometric layout of the scene, our approach infers and generates a large variety of geometrically and photometrically accurate potential images of synthetic pedestrians along with purely accurate ground-truth labels through the use of computer graphics rendering engine. We first present an efficient discriminative learning method that takes these synthetic renders and generates a unique spatially-varying and geometry-preserving pedestrian appearance classifier customized for every possible location in the scene. In order to extend our approach to multi-task learning for further analysis (*i.e.*, estimating pose and segmentation of pedestrians besides detection), we build a more generalized model employing a fully convolutional neural network architecture for multi-task learning leveraging the "free" ground-truth annotations that can be obtained from our pedestrian synthesizer. We demonstrate that when real human annotated data is scarce or non-existent, our data generation strategy can provide an excellent solution for an array of tasks for human activity analysis including detection, pose estimation and segmentation. Experimental results show that our approach (1) outperforms classical models and hybrid synthetic-real models, (2) outperforms various combinations of off-the-shelf state-of-the-art pedestrian detectors and pose estimators that are trained on real data, and (3) surprisingly, our method using purely synthetic data is able to outperform models trained on real scene-specific data when data is limited.

Keywords Training with Synthetic Data · Pedestrian Detection · Pose Estimation

1 Introduction

Over the past decade, computer vision has seen great strides across a wide array of tasks including object recognition and detection [1], semantic segmentation [2], image captioning [3], face recognition [4] and many more. The success of these models depends heavily on the availability of computational resources and a key ingredient for learning such complex models – large amounts of human annotated data. In many scenarios, however, human labeled data is scarce or worse yet, simply unavailable.

In this work we consider one such scenario where a surveillance system is newly installed in a novel location and we must bootstrap a pedestrian detector and pose

* These authors contributed equally.

Hironori Hattori
Institute of Industrial Science, The University of Tokyo
E-mail: hattorih@iis.u-tokyo.ac.jp

Namhoon Lee
Engineering Science Department, University of Oxford
E-mail: namhoon.lee@eng.ox.ac.uk

Vishnu Naresh Boddeti
Computer Science and Engineering,
Michigan State University
E-mail: vishnu@msu.edu

Fares Beainy
Volvo Construction Equipment
E-mail: fares.beainy@volvo.com

Kris M. Kitani, Takeo Kanade
The Robotics Institute, Carnegie Mellon University
E-mail: kkitani@cs.cmu.edu

estimator for a specific surveillance environment without access to any pre-acquired real pedestrian data, either labeled or unlabeled. A similar situation may arise when a new imaging system (*e.g.*, a custom camera with unique lens distortion or internal building surveillance system) has been designed and must be able to detect pedestrian without the expensive and burdensome process of collecting data with the new imaging device or system.

A straightforward solution would be to use an existing generic pedestrian detection and pose estimation system. Most of these generic systems however, are trained on a data distribution that is potentially quite different from the scene under consideration and may result in a very low accuracy system. Although it would be possible to adapt generic models to the new environment by incrementally labeling real examples, it would still require significant manual human intervention. In contrast, in this work we would like to completely automate this process and learn a pedestrian detector and pose estimator without using any real data.

Fortunately, in the aforementioned scenarios, we typically have access to two important pieces of information: (1) the camera's calibration parameters, and (2) scene geometry. With this information, we show that it is possible to generate synthetic training data (*i.e.*, computer generated pedestrian images) to act as a proxy for the real data and to perform human activity analysis such as localizing pedestrians and estimating their pose. Moreover, we show that by using this 'data-free' technique (*i.e.*, does not require real pedestrian data), we are still able to train a scene-specific pedestrian detector that outperforms many baseline techniques.

Using geometrically consistent synthesis of humans presents us with many advantages that can compensate for the lack of real training data: (1) We can maximize the physical geometric information in the scene in terms of the appearance of humans in the scene, the static objects in the scene causing occlusions, resolution and quality of human appearance captured by the camera system, distortions caused by camera optics and partial people at the edges of the camera frame. This geometric information can be incorporated into the data synthesis pipeline to generate realistic renders of virtual humans; (2) We can potentially synthesize an unlimited amount of pedestrian samples spanning a wide range of appearance variations (*e.g.*, clothing, height, weight, gender, ethnicity) on demand; (3) We can simulate human appearance at literally all potential locations in the scene that humans can exist. Additionally we can precisely control the pose, orientation and 3D location of the simulated pedestrian in the scene; (4) We can automatically obtain annotations for detection, body part locations and segmentation masks. The annotations obtained this way are noiseless and precise while human-labeled data is often noisy and error prone.

In our proposed approach, we simultaneously learn hundreds of pedestrian detectors for a single scene using millions of synthetic pedestrian images. Since our approach is purely dependent on synthetic data, the algorithm requires no real-world data. To learn the set of scene-specific location-specific pedestrian detectors, we propose an efficient and scalable appearance learning framework. Our algorithmic framework makes use of highly-efficient correlation filters as our basic detection unit and globally optimizes each model by taking into the account the appearance of a pedestrian over a small spatial region. We compare our approach to several generically trained baseline models and show that our approach generates a better performing scene-specific pedestrian detector. More importantly, our experimental results over multiple data sets show that our 'data-free' approach actually outperforms models that are trained on real scene-specific pedestrian data when data is limited.

We go further and propose to learn a scene-and-region specific spatially-varying fully convolutional neural network for simultaneous detection, pose estimation and segmentation of pedestrians. Traditionally synthetic data has often been used in conjunction with real data while training, either for learning models from scratch or for fine-tuning an existing model. In contrast, the proposed network model is trained purely on synthetic data from scratch. Surprisingly, our method outperforms competitive alternatives that are trained on real data, when evaluated not only on synthetic images but on real data as well, contradicting conventional thought, that models trained purely on synthetic data cannot obtain high accuracy on real data.

The paper is organized as follows: First, we review related approaches in the area of synthetic data, pedestrian detection and pose estimation, and visual analysis for surveillance system in Section 2. Second, we describe our proposed approach about data synthesis in Section 3.1, discriminative learning method in Section 3.2, and fully convolutional neural network architecture for multi-task learning method in Section 3.3. Finally experimental results are described in Section 4.

2 Related Work

2.1 Employing Synthetic Data in Computer Vision

The idea of using synthetic data for 2D object detection is not new. Brooks [5] used computerized 3D primitives to describe 2D images. Dhome *et al.* used computer generated models to recognize articulated objects from a single image [6]. 3D computer graphics models have been used for modeling human shape [7, 8], body-part gradient templates [9], full-body gradient templates [10] and hand appearance [11, 12, 13]. In addition to modeling people, 3D simulation has been used for multi-view car detection [14, 15, 16] and 3D indoor scene understanding [17, 18]. Sun and Saenko [19] used virtual objects to train 2D object detector for real objects. Work by Marin *et al.* [10] used a video game rendering engine to generate synthetic training data. While they learned a single pedestrian detector applied to a mobile scenario, we learn hundreds of location sensitive models for a surveillance scenario. Synthetic data can also be used for evaluation [20]. The main benefit of computer generated data is that it does not require manual data labeling since the ground truth is known. The second benefit is that large amounts of data can be generated with little effort. We take advantage of both of these benefits in our work.

The use of synthetic models has been explored for a variety of computer vision tasks, typically in the context of data augmentation or domain adaptation for object classification. Aubry *et al.* [21] posed object detection as a 2D-3D alignment problem and learned exemplar classifiers from 3D models to align and retrieve the models that best matches the viewpoint of 2D objects in images. Vazquez *et al.* [22] combined synthetic pedestrian data with real pedestrian data to generate robust real world detectors. Pishchulin *et al.* [23] generated pedestrian samples with realistic appearance and backgrounds while modifying body shape and pose using 3D models to augment their real training data for pose estimation. More recently, there are some works using large scale synthetic dataset for video analysis from in-car camera or urban scenes [24, 25, 26]. These techniques demonstrated that the performance of visual classifiers can be improved by augmenting real data with a large amount of synthetic data. We emphasize here that we operate in a different regime where *no real data is available* for augmentation or adaptation.

One effective use of computer generated images is in the area of visual domain adaptation [27, 28, 22]. First, large repositories of synthetic 2D data can be used to bootstrap detectors. Then, the data or detectors can be adapted to real data by leveraging data from the test distribution. Pishchulin *et al.* combined synthesized real 3D human body models and a small number of labeled pedestrian bounding boxes to learn a very robust pedestrian detector [28]. Their work showed that augmenting the training set with the appropriate mix of synthetic and real data can maximize test time performance. Vazquez *et al.* [22] also showed how synthetic pedestrian data can be combined with real pedestrian data to generate robust real-world detectors. We address a different task than domain adaptation, in that we are learning the synthetic pedestrian model needed prior to the adaptation task.

Adapting pre-trained models to a new domain has been an active area of research [29, 30, 31, 32, 33]. The most recent approaches can adapt detectors trained in another domain without the need for new labeled data by bootstrapping a new detector with high or low confidence detections in the new scene [34]. Our work is distinct from work on domain adaption in that we do not allow access to scene-specific real data. In domain adaptation a pre-existing pedestrian detector (or generic pedestrian data) is *augmented* with scene-specific real data to improve performance. Our work is complementary to domain adaptation work in that our proposed detector or data can be used as an initialization for the domain adaptation problem.

Visual analysis tasks can also be trained using only synthetic data [35]. Recently Su *et al.* [36] proposed to use a large collection of 3D models for viewpoint estimation in images. Fischer *et al.* [37] used rendered data of flying chairs for supervised optical flow prediction. However, in these tasks the rendering is done without considering any scene information which results in physically implausible synthetic images (*e.g.*, floating cars). Shotton *et al.* [38] leveraged prior knowledge that the camera will be roughly fronto-parallel to the user to generate a variety of synthetic depth maps to train a human pose estimator. Hattori *et al.* [39] used prior information about the scene to learn scene-specific pedestrian detectors purely from synthetic data. The work showed that leveraging prior camera and scene knowledge in the synthetic data generation pipeline can help to ensure a tighter coupling between people observed in the training and testing data distributions. Our approach builds on the work of [39] but extends to a far more challenging task, *i.e.*, simultaneous articulated human pose estimation and body segmentation in addition to detection. Furthermore, our proposed model is based on a deep convolutional neural network that is trainable end-to-end instead of using a support vector machine on top of hand-crafted features.

2.2 Pedestrian Detection and Human Pose Estimation

There is a large body of work for pedestrian detection and human pose estimation. A complete treatment of this vast literature is beyond the scope of this paper. We instead provide a brief overview of the main techniques and focus on the most relevant state-of-the-art methods. Research on pedestrian detection is largely focused on designing better feature representations and part-based architectures. Carefully designed features [40, 41, 42, 43] that are computationally efficient have been the focus of much of the last decade. For this work, we first limit our choice to the standard HOG feature. For the classifier, we utilize a correlation filter based approach [44, 45] over the standard SVM for computational efficiency reasons as we are required to learn large numbers of templates for a scene.

In contrast, modern day methods for pedestrian detection are based on carefully designed deep network architectures for feature learning [46, 47]. Architecturally, deformable part based methods [48, 49] have been the dominant method for detecting pedestrians. More recently, it has been shown that general object detection frameworks [50, 51, 52] can also achieve competitive pedestrian detection performance. Later in this work, we present a fully end-to-end classifier based on convolutional neural network to enhance the feature engineering stage, and more importantly, to realize multi-task learning for further analysis required in surveillance scenarios.

Interestingly, techniques for human pose estimation have been developed independently from human detection, where it is often assumed that the rough location of the person is available prior to pose estimation. Techniques for human pose estimation can be largely categorized into deformable parts based models [53, 54, 55, 56], deep convolutional networks that regress from the image to the keypoint locations [57, 58, 59] and methods that regress from the image to the ideal localization heat-maps [60, 61, 62, 63] of body parts. Toshev *et al.* [57] introduced one of the earliest deep learning based approaches for pose estimation, learning a regression function from the image to the part coordinates. Carreira *et al.* [58] introduced a similar approach that iteratively refines the prediction of part locations. Current state-of-the-art approaches for human pose estimation, Convolutional Pose Machines (CPM) [61] and Stacked Hourglass Networks [62], directly regress part localization heat maps from the input image. These approaches, 1) assume that humans have been detected, at least coarsely, and 2) are trained on real annotated images spanning a range of human pose and appearance.

Our approach, learns a scene-and-region-specific model that integrates (via heat map regression) pedestrian detection, pose estimation and segmentation into a single fully convolutional neural network. And unlike existing approaches, our model is trained purely on **synthetically rendered** pedestrians and evaluated on **real** pedestrian images. By leveraging geometrically accurate renderings of humans in the scene, our approach is able to bridge the gap in appearance between real and synthetic humans and outperforms generic state-of-the-art approaches trained on real data for human detection and pose estimation for a given scene.

2.3 Visual Analysis for Surveillance Systems

In this work, we have limited ourselves to a surveillance scenario where the camera is static and the scene is known. This stands in contrast to the large body of work focused primarily on pedestrian detection from a mobile platform [64, 65, 43, 66]. The mobile scenario describes a more challenging problem where the camera is undergoing ego-motion and the scene geometry is usually unknown. Adaptive techniques have been proposed for surveillance scenarios [67, 68, 69, 70]. The use of scene geometry, changes in background over time and locality aware detectors can be used to greatly improve the performance of detectors for a specific scene. Our work is similar in that we use scene geometry and camera calibration parameters to generate scene-specific synthetic data. Our work is different in that we do not use real data from the scene to adapt our detector.

3 Proposed Method

Fig. 1 gives a pictorial illustration of our spatially-varying scene-specific pedestrian detection framework. We consider the surveillance setting where the following information is known *a priori.* (however, automated ways of obtaining this information exist): (1) intrinsic and extrinsic parameters of the static camera and (2) the geometrical layout of the scene, *i.e.*, semantic labels for all the regions ("pedestrian region") in the scene where a pedestrian could possibly appear and semantic labels for obstacles in the scene where a pedestrian could either be occluded or physically cannot be present. We call this information scene description. The scene description is leveraged along with synthesized 3D pedestrian models to generate realistic simulations of the appearance of pedestrians for every location of the "pedestrian region". We then learn a smooth spatially-varying scene-specific discriminative appearance model for pedestrian detection. During detection, unlike the conventional approach where a single global detector is applied across the entire image, hundreds of scene-specific location-specific pedestrian detectors are applied to the scene.

In order to scale to multi-task learning for further analysis (*i.e.*, estimating pose and segmentation of pedestrians besides detection), we build a more generalized model employing a fully convolutional neural network

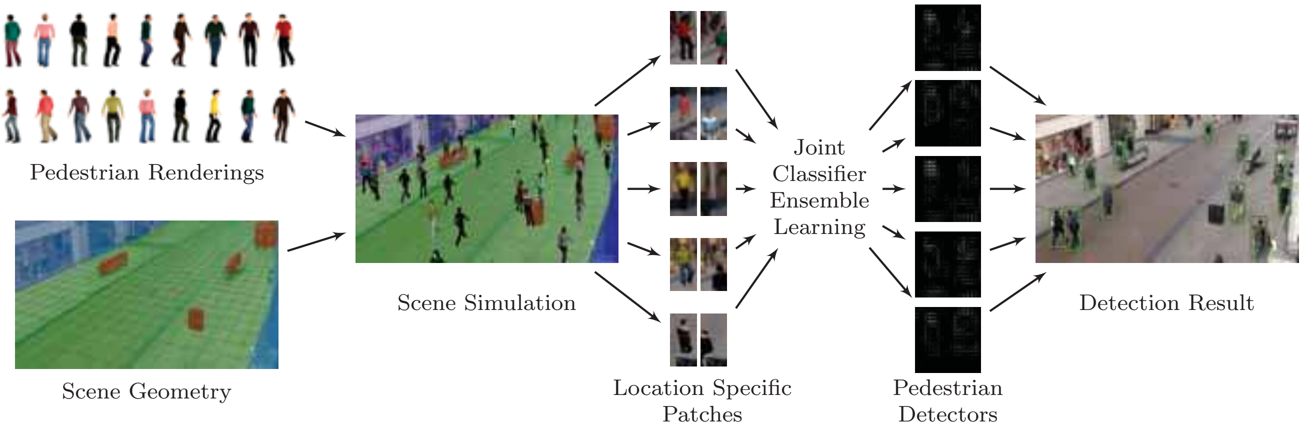


Fig. 1: **Overview of our efficient discriminative learning method:** Given a single image of a scene, camera parameters and coarse scene geometry (*i.e.*, obstacles (red), walls (blue) and walkable areas (green)) as input, our approach synthesizes physically grounded and geometrically accurate renders of pedestrians for every grid location. All location-specific pedestrian detectors are trained jointly to learn a smoothly varying appearance model. Multiple scene-and-location-specific detectors are run in parallel at every grid location.

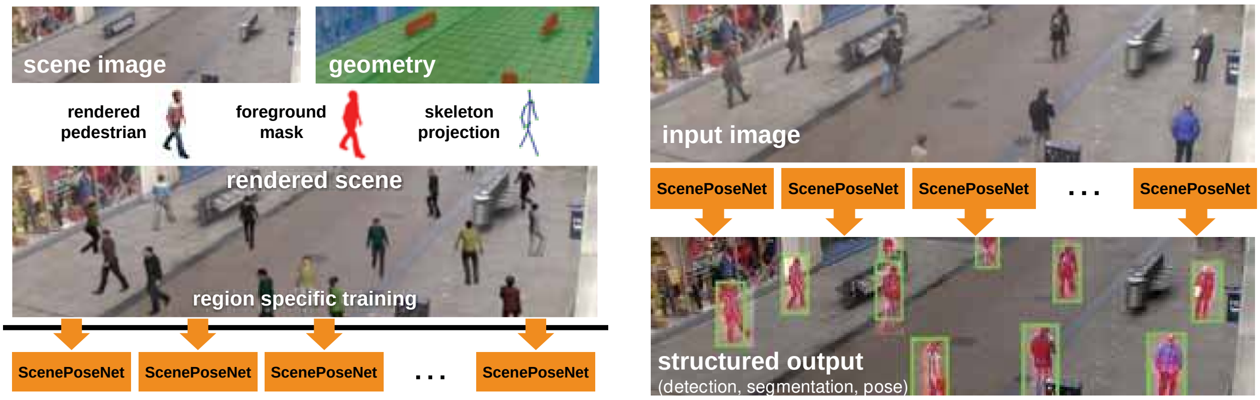


Fig. 2: **Overview of our fully convolutional neural network architecture for multi-task learning method:** With physically grounded and geometrically accurate renders of pedestrians for every grid location, our region specific pedestrian detection and pose estimation networks are trained on this synthetic data. At test time, our model takes a single image and outputs pedestrian detections, segmentation mask and body pose estimates.

architecture for multi-task learning. Fig. 2 gives a pictorial illustration of the inner workings of our approach to generate a scene-specific human detection and pose estimation framework given the scene description. Along with a single image, this scene description serves as the input to our framework to synthesize physically grounded and geometrically accurate humans in the valid regions of the scene. Our approach then learns an ensemble of region-specific models for simultaneous detection, pose estimation and segmentation of humans. During inference, each of these region-specific models are run in parallel on their corresponding regions.

In the following section, we firstly describe the data synthesis approach from scene description (in Section 3.1). Secondly, we describe our discriminative learning method for the scene-specific pedestrian detection framework (in Section 3.2) and our fully convolutional neural network architecture for multi-task learning method for the human detection and pose estimation framework (in Section 3.3).

3.1 Data Synthesis from Scene Description

High quality ground truth annotations are required to train pedestrian detection and pose estimation systems. Obtaining these labels from real data usually requires a costly and noisy process of manual human labeling, a process that does not scale very well to a large number of scenes. Instead, our approach uses the scene description to simulate probable pedestrian appearance appropriate for each region of the scene.

Given the scene description, our approach first generates a planar 3D model of the scene, *i.e.*, fits a planar ground plane, planar walls and cuboids to encompass the obstacles. The camera parameters can then be used to account for camera lens characteristics (*e.g.*, perspective distortion in wide-angle cameras) and the scene viewpoint for rendering geometrically accurate humans. Autodesk 3DS Max is used as the scene modeling and rendering engine. The rendering pipeline can precisely control the following variations in human appearance: gender, height, width, orientation and pose in addition to rendering human appearance at every valid pedestrian location of the scene. Our virtual human database consists of 139 different models spanning gender, clothing color and skin color. The models used for this work only have skin tight clothing but have a continuous range of walking configurations from standing to running. Our approach uniformly samples body orientations from 0° $\sim$ 360° (Fig. 3) but can also be guided by any prior information if available.

To generate ground truth labels for the people in the rendered images we first associate attributes to each 3D virtual model with the following labels: segmentation mask, 3D locations of 27 parts and the location of the center of the person for detection. The 2D labels for training can then be automatically extracted from the 3D annotations and the camera projection parameters. This process allows us to generate consistent noise free labels, unlike human annotations, at scale across all rendered images. Furthermore, we can also uniformly span all the variations in appearance, orientation, pose or location unlike real data that follows a long-tailed distribution.

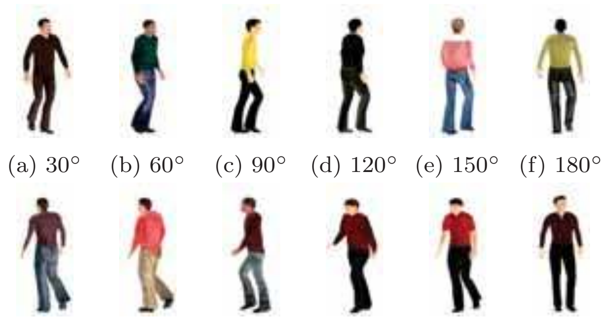

(a) 30° (b) 60° (c) 90° (d) 120° (e) 150° (f) 180°
(g) 210° (h) 240° (i) 270° (j) 300° (k) 330° (l) 360°

Fig. 3: A few examples of the pedestrian renderings used for training our pedestrian detectors. We have a total of 36 different pedestrian models and for each location we simulate pedestrians with 3 different walking poses and 12 (every 30°) different orientations.

3.2 Discriminative Learning Method

In this section we describe how to realize discriminative learning with the data sythesized in Section 3.1.

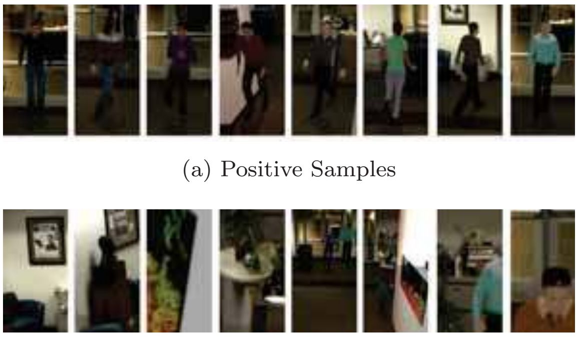
3.2.1 Classifier Ensemble Learning

During detection using our approach, unlike the conventional approach where a single global detector is applied across the entire image, hundreds of scene-specific location-specific pedestrian detectors are applied to the scene. Those detectors are trained for each location of a fixed grid of single scale square regions (like 8 × 8 or 16 × 16). Since the detectors at each location in the image significantly overlap with each other it is natural to impose smoothness constraints between neighboring detectors – neighboring detectors should be similar. Therefore, we propose a joint detector learning approach while imposing smoothness constraints between neighboring detectors. A nice consequence of our framework is that the detectors are implicitly calibrated since they are jointly trained. We base our detector on the Vector Correlation Filter [44] formulation where the detector design is posed as a regression problem.

Notation: For notational ease, all expressions through the rest of this paper are given for 1-D signals with K-channels. Vectors are denoted by lower-case bold ($\mathbf{x}$) and matrices in upper-case bold ($\mathbf{X}$). $\hat{\mathbf{x}} \leftarrow \mathcal{F}_K(\mathbf{x})$ and $\mathbf{x} \leftarrow \mathcal{F}_K^{-1}(\hat{\mathbf{x}})$ denotes the Fourier transform of $\mathbf{x}$ and the inverse Fourier transform of $\hat{\mathbf{x}}$, respectively, where $\hat{}$ denotes variables in the frequency domain, $\mathcal{F}_K()$ is the Fourier transform operator and $\mathcal{F}_K^{-1}()$ is the inverse Fourier transform operator with the operators acting on each of the K channels independently. Superscript † denotes the complex conjugate transpose operation.

We pose the problem of jointly learning the n detectors with m_i training samples per detector as the following optimization problem:

$$\min_{\mathbf{w_1},\ldots,\mathbf{w_n}} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left\|\sum_{k=1}^{K}\mathbf{x}_i^{kj} * \mathbf{w}_i^k - \mathbf{g}_i^j\right\|_2^2 + \frac{\lambda}{2}\sum_{(i,j)\in E} c_{ij}\|\mathbf{w}_i - \mathbf{w}_j\|_2 \qquad (1)$$



(a) Positive Samples

(b) Negative Samples

Fig. 4: The positive samples have variations in pedestrian pose, appearance, height, gender etc. On the other hand the negative samples consist of many variations of the background including samples with partial occluded pedestrians and pedestrians at very different scale.

where $*$ denotes the correlation operation, $\mathbf{x}_i^j$ is the j-th training image for the i-th detector $\mathbf{w}_i$, $\mathbf{g}_i^j$ is the desired correlation response and E defines the set of edges[1] with connections between neighboring regions that overlap with each other. The second term captures the classifier smoothness constraints for overlapping regions of the classifier and c_{ij} captures the smoothness weights and λ is the regularization parameter which trades-off the smoothness term. We adopt the Alternating Direction Method of Multipliers (ADMMs) [71] to solve the above optimization problem efficiently. The problem is now posed as:

$$\begin{aligned}\min_{\mathbf{w_1},\dots,\mathbf{w_n}} \quad & \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left\|\sum_{k=1}^{K}\mathbf{x}_i^{kj} * \mathbf{w}_i^k - \mathbf{g}_i^j\right\|_2^2 \\ & + \frac{\lambda}{2}\sum_{(i,j)\in E} c_{ij}\|\mathbf{h}_i - \mathbf{h}_j\|_2^2 + \frac{\rho}{2}\|\mathbf{W}-\mathbf{H}\|_F^2 \\ s.t. \quad & \mathbf{W} = \mathbf{H}\end{aligned} \tag{2}$$

where ρ is a regularization parameter. We now form and optimize the Lagrangian for this optimization problem,

$$\begin{aligned}L(\mathbf{W},\mathbf{H},\boldsymbol{\Lambda}) = & \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left\|\sum_{k=1}^{K}\mathbf{x}_i^{kj} * \mathbf{w}_i^k - \mathbf{g}_i^j\right\|_2^2 \\ & + \frac{\lambda}{2}\sum_{(i,j)\in E} c_{ij}\|\mathbf{h}_i - \mathbf{h}_j\|_2^2 \\ & + \frac{\rho}{2}\|\mathbf{W}-\mathbf{H}\|_F^2 \\ & + \Lambda^T(vec(\mathbf{W}) - vec(\mathbf{H}))\end{aligned} \tag{3}$$

This problem can be solved by decomposing it into subproblems for $\mathbf{W}$, $\mathbf{H}$ and $\boldsymbol{\Lambda}$, each of which can in turn be solved very efficiently.

Subproblem W:

$$\mathbf{W}^{l+1} = \arg\min_{\mathbf{W}} L(\mathbf{W},\mathbf{H}^l,\boldsymbol{\Lambda}^l) \tag{4}$$

This sub-problem can be further decomposed into individual sub-problems for each of the locations in the scene in closed form in the Fourier domain i.e,

$$\begin{aligned}\mathbf{w}_i^{l+1} &= \arg\min_{\mathbf{w}} \frac{1}{2}\sum_{j=1}^{m_i}\|\sum_{k=1}^{K}\mathbf{x}_i^{kj} * \mathbf{w} - \mathbf{g}_i^j\|_2^2 \\ &\quad + \frac{\rho}{2}\|\mathbf{w} - \mathbf{h}_i^l\|_2^2 + \boldsymbol{\Lambda}_i^{lT}(\mathbf{w} - \mathbf{h}_i^l) \\ &= \mathcal{F}_K^{-1}\{\arg\min_{\hat{\mathbf{w}}} \frac{1}{2}\sum_{j=1}^{m_i}\|\sum_{k=1}^{K}\hat{\mathbf{X}}_i^{kj\dagger}\hat{\mathbf{w}} - \hat{\mathbf{g}}_i^j\|_2^2 \\ &\quad + \frac{\rho}{2}\|\hat{\mathbf{w}} - \hat{\mathbf{h}}_i^l\|_2^2 + \hat{\boldsymbol{\Lambda}}_i^{l\dagger}(\hat{\mathbf{w}} - \hat{\mathbf{h}}_i^l)\} \\ &= \mathcal{F}_K^{-1}\{(\hat{\mathbf{D}} + \rho\mathbf{I})^{-1}(\rho\hat{\mathbf{h}}_i^l + \hat{\mathbf{p}} - \hat{\boldsymbol{\Lambda}}_i^l)\}\end{aligned}$$

where we use the Parseval's theorem to express the objective function in the Fourier domain. $\hat{\mathbf{X}}_j^{k\dagger}\hat{\mathbf{h}}^k$ is the DFT (of size $N_{\mathcal{F}}$) of the correlation of the k-th channel of the j-th training image with the corresponding k-th channel of the CF template where the diagonal matrix $\hat{\mathbf{X}}_j^k$ contains the vector $\hat{\mathbf{x}}_j^k$ along its diagonal and,

$$\hat{\mathbf{D}} = \frac{1}{N_{\mathcal{F}}}\begin{bmatrix}\sum_{j=1}^{m_i}\hat{\mathbf{X}}_j^1\hat{\mathbf{X}}_j^{1\dagger} & \cdots & \sum_{j=1}^{m_i}\hat{\mathbf{X}}_j^1\hat{\mathbf{X}}_j^{K\dagger} \\ \vdots & \ddots & \vdots \\ \sum_{j=1}^{m_i}\hat{\mathbf{X}}_j^K\hat{\mathbf{X}}_j^{1\dagger} & \cdots & \sum_{j=1}^{m_i}\hat{\mathbf{X}}_j^K\hat{\mathbf{X}}_j^{K\dagger}\end{bmatrix} \tag{5}$$

$$\hat{\mathbf{p}} = \frac{1}{N_{\mathcal{F}}}\begin{bmatrix}\sum_{l=1}^{m_i}\hat{\mathbf{X}}_i^{1j}\hat{\mathbf{g}}_i^j \\ \vdots \\ \sum_{l=1}^{m_i}\hat{\mathbf{X}}_i^{Kj}\hat{\mathbf{g}}_i^j\end{bmatrix}, \hat{\mathbf{h}} = \begin{bmatrix}\hat{\mathbf{h}}^1 \\ \vdots \\ \hat{\mathbf{h}}^K\end{bmatrix}, \hat{\mathbf{w}} = \begin{bmatrix}\hat{\mathbf{w}}^1 \\ \vdots \\ \hat{\mathbf{w}}^K\end{bmatrix}$$

Subproblem H:

$$\mathbf{H}^{l+1} = \arg\min_{\mathbf{H}} L(\mathbf{W}^{l+1},\mathbf{H},\boldsymbol{\Lambda}^l) \tag{6}$$

The solution for this sub-problem results in a closed form solution which can be implemented very efficiently in the spatial domain.

$$\begin{aligned}\mathbf{H}^{l+1} &= \arg\min_{\mathbf{H}} \frac{\lambda}{2}\sum_{(i,j)\in E} c_{ij}\|\mathbf{h}_i - \mathbf{h}_j\|_2^2 \\ &\quad + \frac{\rho}{2}\|\mathbf{W}^l - \mathbf{H}\|_F^2 + \boldsymbol{\Lambda}^{lT}(vec(\mathbf{W}^l) - vec(\mathbf{H})) \\ &= \arg\min_{\mathbf{H}} \frac{\lambda}{2}\mathbf{H}^T\mathbf{A}\mathbf{H} + \frac{\rho}{2}\|\mathbf{W}^l - \mathbf{H}\|_F^2 \\ &\quad + \boldsymbol{\Lambda}^{lT}(vec(\mathbf{W}^l) - vec(\mathbf{H})) \\ &= (\lambda\mathbf{A} + \rho\mathbf{I})^{-1}(\rho vec(\mathbf{W}) + vec(\boldsymbol{\Lambda}))\end{aligned}$$

[1] The connectivity graph considered in this paper is a Markov Random Field over all regions while ignoring the regions defined as walls and obstacles.

where $\mathbf{A}$ is a sparse adjacency matrix defining the connectivity structure (defined by the scene geometry) of the smoothness graph.

Subproblem $\mathbf{\Lambda}$:

$$\mathbf{\Lambda}^{l+1} = \mathbf{\Lambda}^{l} + \rho(\mathbf{W}^{l+1} - \mathbf{H}^{l+1}) \tag{7}$$

3.2.2 Detection Protocol

Given a video frame, pedestrian detection is performed by running each of the spatially varying pedestrian detectors at their corresponding locations resulting in a detection response map over the entire image. To account for the height variation among pedestrians we also evaluate the detectors over a small range of scales at each location (0.95 to 1.05). Finally we apply non-maximal suppression to filter the multiple overlapping detections for each instance of the object obtained from the response map. We note that due to the spatially varying nature of the pedestrian detector, detection can no longer be performed efficiently using convolution.

3.3 Fully Convolutional Neural Network Architecture for Multi-task Learning Method

In this section we describe how to realize fully convolutional newral network architecture for multi-task learning with the data synthesis in Section 3.1.

3.3.1 Learning the Network from Synthetic Data

Using the scene-specific data generated above, our now develop a deep neural network trained to operate according to the specifications of the scene description. Our deep convolutional neural network improves over our initial model by providing a more general architecture for multi-task learning applied to pedestrian detection.

The network trained by our approach is designed to jointly accomplish the following tasks: localization of pedestrians, localizing the landmarks that define their pose and segment the pixels that define them. To predict the pedestrian location, pose and segmentation mask the network has to model the full appearance of the pedestrian, the local appearance of the landmarks and a prior on the valid spatial configuration of these parts. The network design aims to encapsulate these desiderata. To capture appearance, both the full pedestrian and the local landmark appearance, learning is posed as a regression problem mapping the RGB input into a heatmap for accurate localization of the pedestrian, local landmarks and the segmentation mask. The prior on the spatial relationships between the part locations is implicitly learned through a spatial belief module that accounts for the correlations between the heatmaps of the full pedestrian, local landmarks and the segmentation masks. We call our specific instantiation of multi-task deep learning for pedestrian analysis as ScenePoseNet.

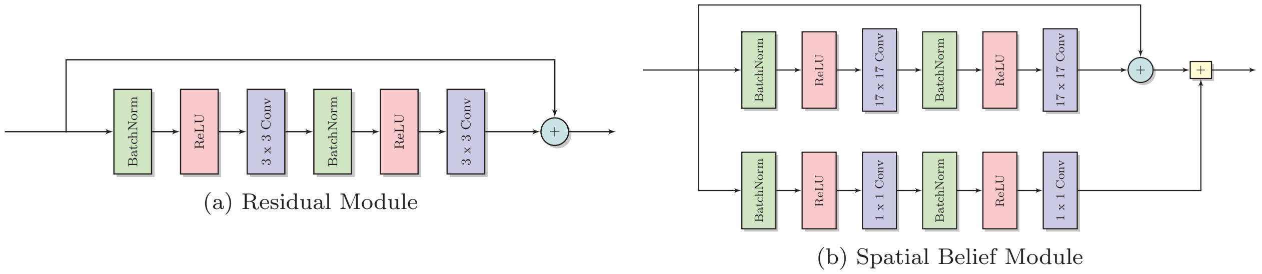
Human pose estimation systems typically treat detection and pose estimation as independent and sequential tasks, with detection followed by pose estimation. These systems either expect ground truth human detections or at least expect a coarse detection using an off-the-shelf detector. However, the tasks of detection and part localization are highly interdependent processes. The accuracy of the detection can greatly affect the pose estimation process. Accordingly, the ScenePoseNet model couples these tasks to improve the efficacy of both pedestrian detection and pose estimation. The main idea behind our ScenePoseNet architecture is to (1) directly regress part localization beliefs from the image features and (2) learn the interactions between these confidence maps.

3.3.2 Basic blocks

We use the following basic units to define our network: Residual Unit [72] and Spatial-Belief Module. The residual unit was introduced to address the problem of vanishing gradients in training very deep convolutional networks. We adopt this basic unit for our network and also build upon it to define our Spatial-Belief (SB) module. As shown in Fig. 5(b) the SB module is purposed to (1) map the input features of the block to the part localization beliefs (heat-maps) and simultaneously (2) process the input features and part localization beliefs from the previous block. The image features and the part localization beliefs generated by this block are concatenated to form the input to the next block. Given an input $\mathbf{x}$ to SB module, the output $\mathbf{y}$ is given by,

$$\begin{aligned} \mathbf{y} &= (\mathbf{x} + f_{res}(\mathbf{x})) \boxplus f_{belief}(\mathbf{x}) & (8)\\ &= (\mathbf{x} + \mathbf{r}) \boxplus \mathbf{b} & (9) \end{aligned}$$

where $\boxplus$ denotes the concatenation operation, $\mathbf{r} = f_{res}(\mathbf{x})$ is the operation through the non-identity branch of the residual unit and $\mathbf{b} = f_{belief}(\mathbf{x})$ denotes the mapping from the input $\mathbf{x}$ to the desired heat maps (human detection, part detection and segmentation mask) through a series of 1×1 convolutions. Our SB unit enables the network to consider part detection confidences with varying amounts of contextual information around the parts from different receptive fields. The part localization confidences $\mathbf{b}_i$ from the i-th SB unit propagates to the next $(i+1)$-th SB block and is processed through the non-identity path where the correlations between



(a) Residual Module

(b) Spatial Belief Module

Fig. 5: The basic modules that comprise our ScenePoseNet architecture, (a) Residual module [72], (b) Spatial-Belief (SB) module. The spatial-belief unit aggregates (⊞ denotes the concatenation operation) the image features extracted from the convolutional network and the confidence maps of the full pedestrian, local landmarks and the segmentation mask from the output of the previous SB unit. The aggregated features now serve as the input to the next SB unit where the image features and confidence maps are jointly processed, thereby learning a prior on valid spatial relationships between the heat maps, and consequently body pose.

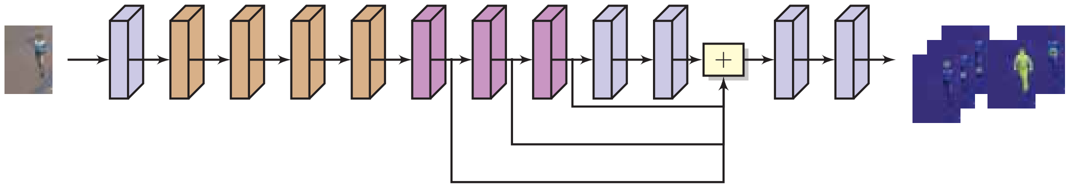


Fig. 6: **ScenePoseNet:** An illustration of our network architecture. Our network is comprised of three basic units: convolutional block, residual block and spatial-belief block. Our network uses information from multiple different spatial contextual regions via skip connections (⊞ denotes the concatenation operation). The input image is mapped to the ideal heat maps for part localization, pedestrian center and segmentation mask.

the heat-maps of the various parts are implicitly captured. This can be readily seen by applying the SB unit operation recursively,

$$\mathbf{x}_{i+1} = (\mathbf{x}_i + \mathbf{r}_i) \boxplus \mathbf{b}_i. \tag{10}$$

Both the identity shortcut and the $f_{res}()$ in each SB unit implicitly processes the beliefs from all previous SB units due to our concatenation operation. Furthermore, the detection confidence maps generated in each SB unit also consider part localization confidences at all previous SB units, each computed with different receptive fields. Therefore, the network takes advantage of detection confidence maps at multiple stages and through multiple receptive field sizes.

3.3.3 ScenePoseNet

Our complete detection, pose estimation and segmentation network architecture is illustrated in Fig. 6. Given an input image, ScenePoseNet jointly localizes pedestrians, localizes body parts and segments the pedestrians in the form of heat maps. The network is composed of fully convolutional layers to preserve spatial context while being computationally efficient. For precise localization and pose estimation of pedestrians we also use dense heat map prediction throughout the network preventing loss of information due to sub-sampling (pooling). The input image is passed through a convolutional layer with 5×5 filters, followed by four residual units with 3×3 filters following the design of residual networks for object recognition. This is followed by 3 SB units each with convolutional filters with large receptive fields, 17×17 to increase the receptive field of the network while still performing dense prediction. The SB units are followed by two 1×1 convolutional layers to map the image features to the heat maps. Finally, skip connections are used for fusing information from multiple different contextual regions, as it combines features from various scales of receptive fields (similar to [61]). The bounding box location for detection is inferred around the heatmaps of joints, center of body, and segmentation.

The network is optimized to minimize the multi-task mean-squared-error loss $\mathcal{L}$ between the network prediction $\{\mathbf{o}_{det}, \mathbf{o}_{pose}, \mathbf{o}_{seg}\} = f_{conv}(\mathbf{b}_i \boxplus \cdots \boxplus \mathbf{b}_n)$ and the ideal heatmaps for pedestrian detection, part localization and segmentation mask, defined as follows,

$$\mathcal{L} = \alpha\mathcal{L}_{det} + \beta\mathcal{L}_{pose} + \gamma\mathcal{L}_{seg} \tag{11}$$

$$\mathcal{L}_{det} = \|\mathbf{o}_{det} - \mathbf{g}_{det}\|_2^2 \quad (12)$$

$$\mathcal{L}_{pose} = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{o}_{pose} - \mathbf{g}_{pose}\|_2^2 \quad (13)$$

$$\mathcal{L}_{seg} = \|\mathbf{o}_{seg} - \mathbf{g}_{seg}\|_2^2 \quad (14)$$

where α, β and γ are hyperparameters trading-off the different loss functions.

4 Experimental Evaluation

In this section, we firstly describe the dataset for experimental results in section 4.1. Then the results by discriminative learning method (Section 4.2) and results by fully convolutional neural network architecture for multi-task learning method (Section 4.3) are follows.

4.1 Datasets

We evaluate the efficacy of our proposed scene-specific spatially-varying pedestrian detection and human pose estimation framework on three different datasets: an outdoor dataset, a semi-outdoor datatset and an indoor dataset.

Towncenter Dataset [73]: The town center dataset is a video dataset of a semi-crowded town center with a resolution of 1920 × 1080 and a frame rate of 25fps. We down-sample the videos to a standardized resolution of 640 × 360.

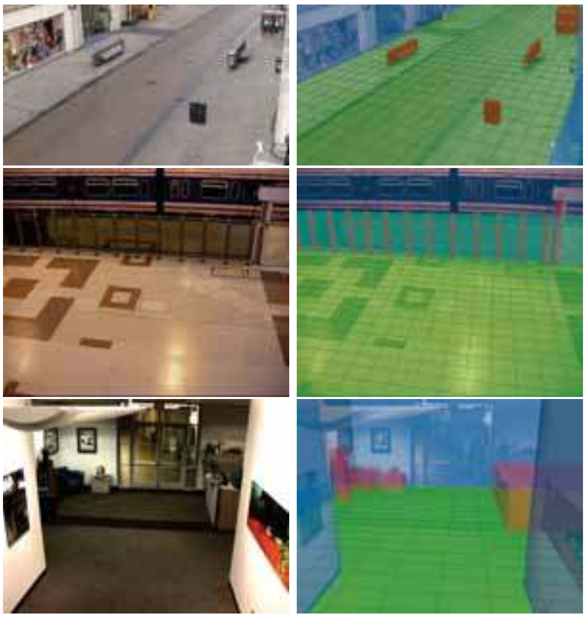


Fig. 7: Three evaluation scenes with their corresponding geometric labels. Town Center [73] (top), PETS 2006 [74] (middle) and CMUSRD [75] (bottom).

PETS 2006 Dataset [74]: The PETS 2006 datatset consists of video (at a resolution of 720 × 576) of a public space including a number of pedestrians. While the dataset consists of videos captured by four different cameras we just use a single camera view for our experiments since our approach is based on a single camera. We down-sample the videos to a standardized resolution of 640 × 512.

CMUSRD [75]: The Carnegie Mellon University Surveillance Research Dataset is a new dataset for indoor surveillance. The data is collected using multiple cameras inside a building at the Carnegie Mellon University and consists of several tens of different people as subjects. While the original resolution of the data is 1280 × 960, we down-sample the videos to a standardized resolution of 640 × 480.

4.2 Discriminative Learning Method

4.2.1 Baselines

We evaluate and compare against the following baseline approaches for the task of pedestrian detection.

G: A single HOG+SVM based pedestrian detector trained on INRIA pedestrian dataset [40].

G+: A single HOG+SVM based pedestrian detector trained on the INRIA pedestrian dataset augmented with negative background patches from the corresponding specific scene.

SS: A single HOG+SVM based pedestrian detector trained on real data from the corresponding specific scene.

DPM: The deformable parts based [76, 77] pedestrian detector trained on the PASCAL VOC *person* class.

DPM+: We build upon the pioneering work by Hoeim et.al.,[78] to leverage the known ground truth scene geometry and camera location/viewpoint at the inference stage using the DPM pedestrian detector as our base detector.

SSV: A single HOG+SVM based pedestrian detector trained **only** on virtual pedestrians whose appearance is simulated in the specific scene under consideration.

SSV+: A single HOG+SVM based pedestrian detector trained on **both** real and virtual pedestrians whose appearance is simulated in the specific scene under consideration. This baseline is similar in spirit to the approach in [22].

SLSV(Ours): Our proposed scene-specific pedestrian detection framework with a spatially varying pedestrian appearance model. This model is learned **entirely** from virtual pedestrians whose appearance is simulated in the specific scene under consideration. In the experiments that follow we train a detector for each 16×16 im-

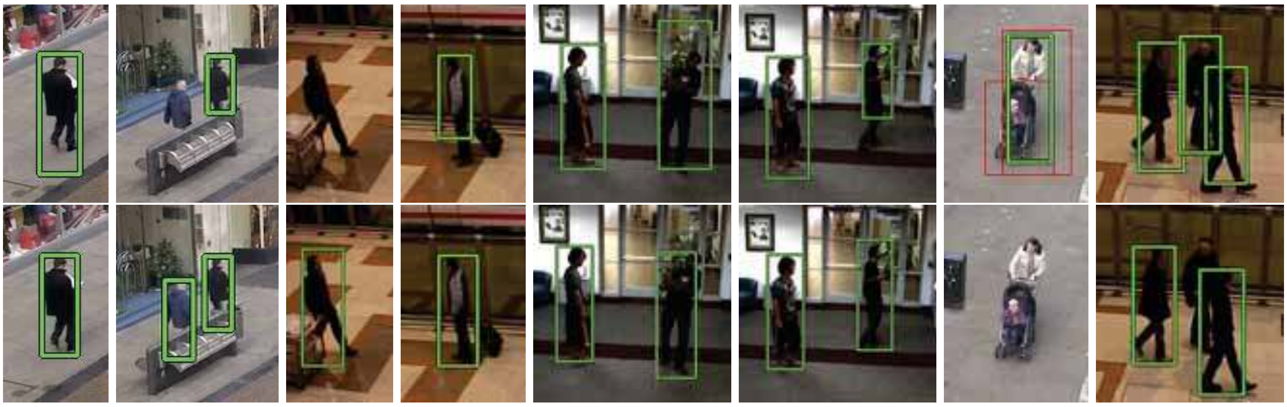


Fig. 8: Sample detections of DPM (**top**) and our proposed method (**bottom**), **green** denotes true positives and **red** denotes false positives.

age patch. The number of models learned for the Town Center, PETS 2006 and CMUSRD is 640, 879 and 348, respectively. Each model is trained using 4000 examples (2000 positive and 2000 negative). This translates to roughly 2.5 million synthetic images used to train the detectors for the Towncenter scene.

4.2.2 2D Bounding Box Evaluation

We compare our proposed model to all baselines using the standard 50% overlap metric used for pedestrian detection [43]. In addition to this metric, we also include results of the 70% overlap criteria to show the 2D localization power of our approach. Results are summarized as PR curves in Fig. 9. The curves show that our approach has a significantly better recall rate due to the ability to learn accurate location specific detectors. The qualitative examples are given in Fig. 8 also illustrate the ability of our method to accurately localize pedestrians. Failure cases also show that our model is not able to detect pedestrians occluded by other pedestrians since this type of occlusion was not generated during training. Table 1 shows the mean average precision (AP) over all three datasets. Our proposed approach using purely synthetic data outperforms all baselines with an AP of 0.90. The DPM+ which uses the same geometry and camera information as our approach performs second best with an AP of 0.86 followed by vanilla DPM with an AP of 0.73. All other models fall closely behind the DPM. The main difference between our approach and the other models is that specific detectors are learned for each location in the scene. Furthermore unlike DPM+ which leverages known scene geometry and camera parameters at inference our model uses the same information at the training stage.

More importantly, we observe that our approach is resilient to a more stringent criteria. Across all three datasets, the standard HOG+SVM model **G** drops by 37% (0.70 → 0.44) and the DPM+ model performance drops by 41% (0.86 → 0.51). In contrast, the performance of the propose method only drop by 22% (0.90 → 0.70) under the tighter criteria. We will examine the localization power of our approach further in section 4.2.5.

Table 1: Average precision by bounding box overlap criteria

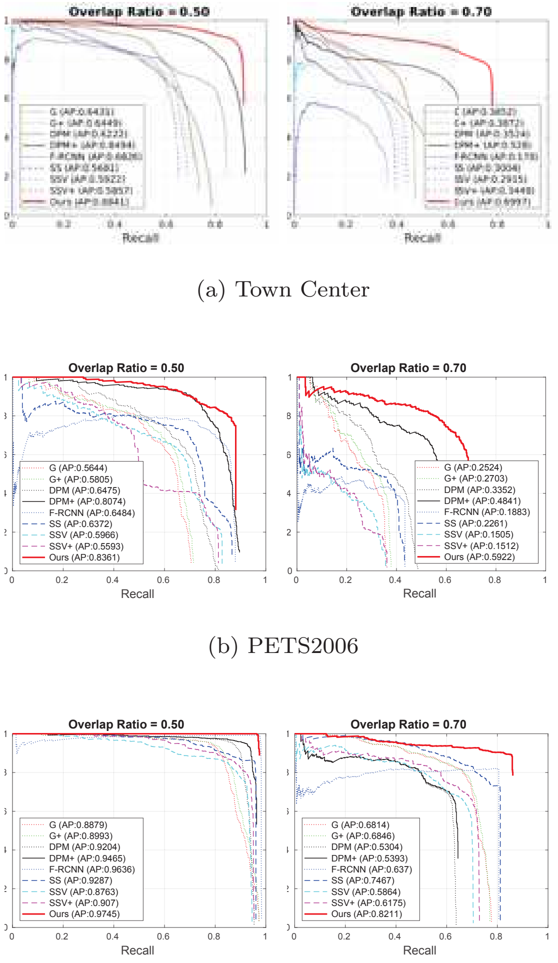
	0.5 overlap	0.7 overlap	Change
G [40]	0.70	0.44	37%
G+	0.71	0.45	37%
DPM [76]	0.73	0.41	44%
DPM+ [78]	0.86	0.51	41%
SS	0.71	0.42	40%
SSV	0.69	0.34	50%
SSV+ [22]	0.68	0.37	46%
SLSV (Ours)	**0.90**	**0.70**	**22%**

4.2.3 Effect of the smoothness constraint

We have formulated a joint learning problem to ensure that appearance models vary smoothly over space. Fig. 10 shows how overall performance is effected by changing the weight of the smoothness term λ in Equation (2). For the CMUSRD dataset, the smoothness constrain improves performance by 8 points (0.89 → 0.97). We obtain optimal performance at a value of $\lambda = 0.10$ for the CMUSRD dataset which we used for all experiments in this paper.

4.2.4 Effect of grid-size resolution

While we observe from comparative experiments that a single generic detector is not flexible enough to cover



(a) Town Center

(b) PETS2006

(c) CMUSRD

Fig. 9: Precision-recall curves in 2D bounding box evaluation for differing overlap ratio criteria.

the entire scene, we would like to understand how many detectors are needed to effectively cover all appearance variations. We evaluated the effect of the grid-size on system performance using a small portion of the Towncenter scene to understand how appearance is affected by location. Table 2 shows how AP performance changes with respect to the grid size (number of learned detectors). The results indicate that a smaller grid size of 8 × 8 patches perform better which means that pedestrian appearance is in fact varying significantly by location. Our results show a plateau effect starting at 16 × 16 so we use this setting for all our experiments.

4.2.5 Localization in 3D

Pedestrian detection is often used as a pre-processing step for tracking, action recognition or activity analysis. In these scenarios, it is helpful to know the precise 3D location of a person in the environment. To evaluate the performance of 3D localization we use a minimum distance metric where a detection is considered valid only if it is within 90 cm of the ground truth location. Table 3 shows mean AP scores over all three dataset. Our proposed approach performs best with a AP of 0.91. The second best is the **SS** model trained on real scene-specific data with an AP of 0.70, followed by other models trained on scene-specific data with **SSV** at an AP of 0.66 and **SSV+** at an AP of 0.65.

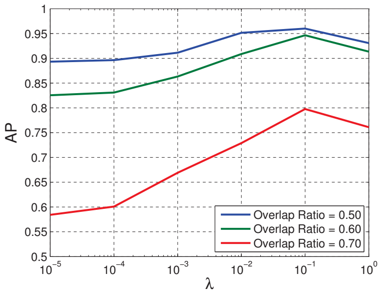


Fig. 10: AP on CMUSRD for different smoothing values (λ).

Table 2: Average precision by number of detectors

Patch Size	Number of Detectors	AP
8 × 8	371	0.802
16 × 16	102	0.798
32 × 32	30	0.764

We also evaluate our approach with a much tighter criteria of 50 cm. Our proposed approach is most resilient to the tighter criteria with a performance drop of only 8% (0.91 → 0.84). The performance of all other models, with the exception of DPM and DPM+, drops between 10% ∼ 13%. The performance of **DPM** and **DPM+** drops by a large 35% and 30% respectively, which indicates that the vertical localization of the DPM model is noisy.

Fig. 11 compares the 3D trajectories of our proposed approach and the DPM model. Bounding box results are projected to the ground plane using the center of the bottom of the box. Since our proposed approach is able to accurately localize pedestrians in the image plane, the projected 3D trajectories are smooth and very close to the ground truth 3D trajectories. The DPM result projected into 3D is quite jagged as the bounding box tends to move up and down during detection.

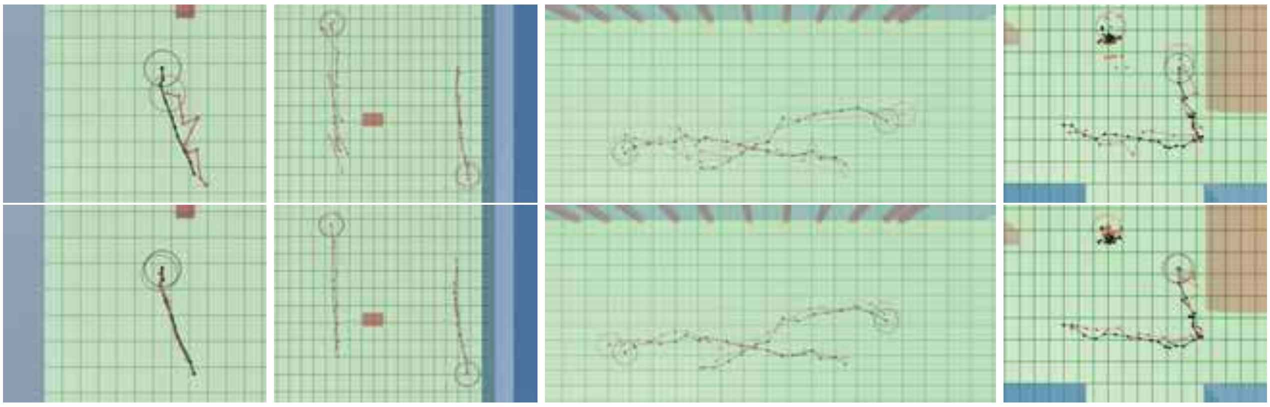
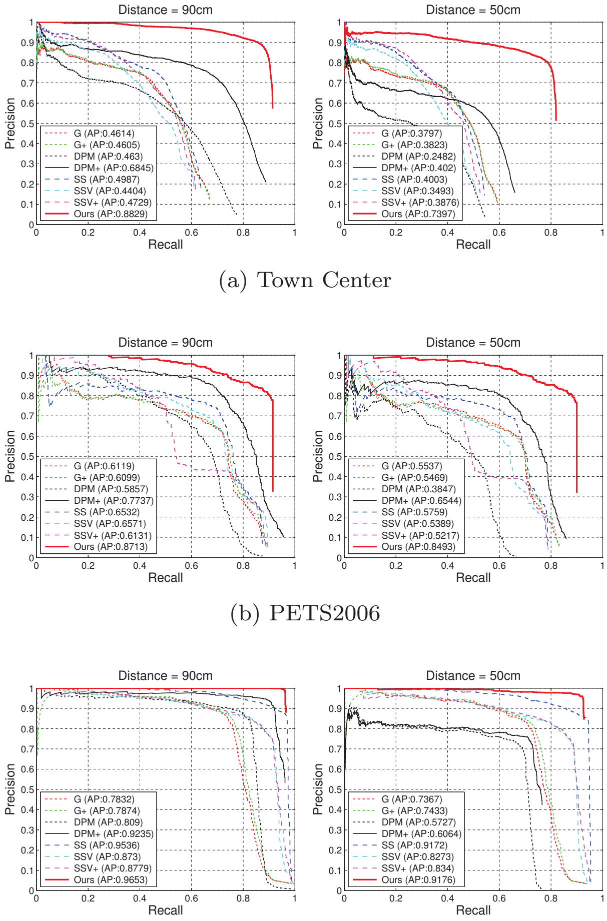


Fig. 11: 3D localization trajectories of DPM (**top**) and our proposed method (**bottom**) in **red**, while the ground truth is in **black**.



(a) Town Center

(b) PETS2006

(c) CMUSRD

Fig. 12: Precision-Recall Curves in 3D localization on Town Center, PETS 2006 and CMUSRD for different amounts of distance.

Table 3: Average precision by 3D distance criteria

	90 cm	50 cm	Change
G [40]	0.62	0.56	10%
G+	0.62	0.56	10%
DPM [76]	0.62	0.40	35%
DPM+ [78]	0.79	0.55	30%
SS	0.70	0.63	10%
SSV	0.66	0.57	13%
SSV+ [22]	0.65	0.58	11%
SLSV (Ours)	**0.91**	**0.84**	**8%**

4.3 Fully Convolutional Neural Network Architecture for Multi-task Learning Method

For a given specific scene we evaluate the efficacy of our approach to generate a ScenePoseNet, for pedestrian detection, pose estimation and segmentation. Detection and pose estimation are evaluated both quantitatively and qualitatively, while segmentation is evaluated only qualitatively due to lack of ground truth segmentation masks. Fig. 13 shows the activation maps at various stages of ScenePoseNet. The spatial belief blocks progressively refine the activation maps from the residual blocks. We note that combining the activation maps from the different spatial belief blocks further improves pedestrian localization in terms of the segmentation mask.

4.3.1 Baselines

We compare our approach based pedestrian detection and pose estimation approach to a number of combinations of state-of-the-art pedestrian detectors and human pose estimation approaches. For pedestrian detection we consider the two baselines that are based on HoG features, SLSV [39] and Deformable Parts Model (DPM) [54], and Faster Region-based Convolutional Neu-

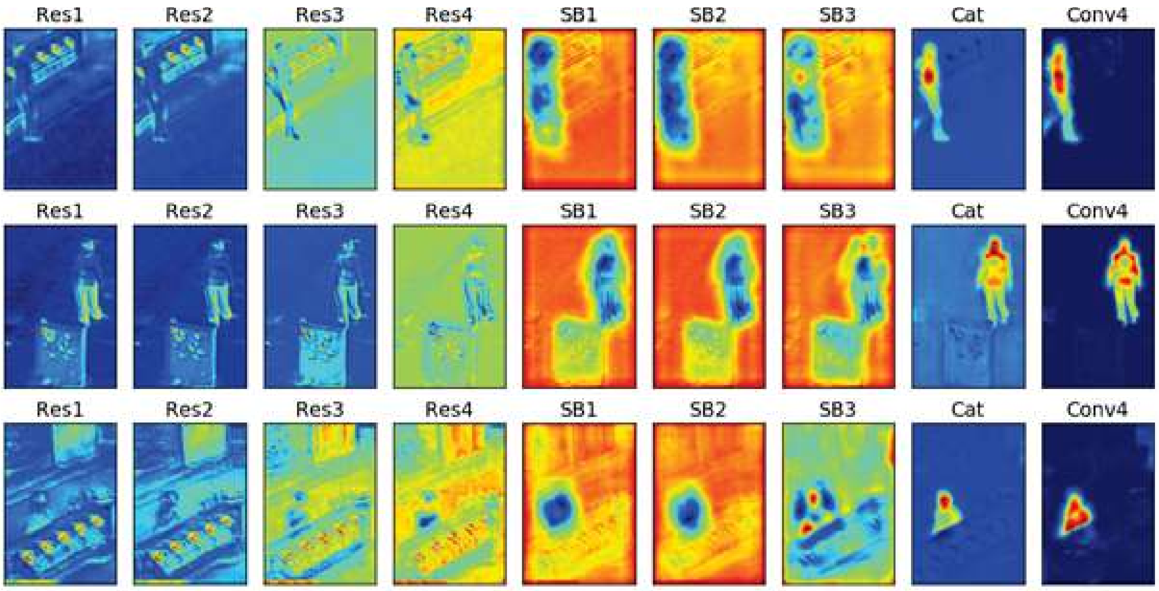


Fig. 13: Visualization of activation map extracted from the intermediate layers of ScenePoseNet for different regions of the scene. As the image propagates through ScenePoseNet, the beliefs of the scene background are suppressed while the beliefs on the pedestrian and the individual joints increases.

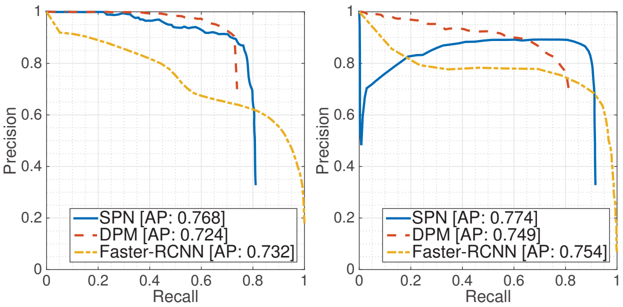


Fig. 14: Precision-Recall curves along with the average precision for pedestrian detection on the (a) Towncenter and (b) PETS2006 datasets. We compare our ScenePoseNet (SPN) with DPM and Faster R-CNN.

ral Network [51], pre-trained on ImageNet and VOC2007. For human pose estimation we compare against two state-of-the-art methods, Convolutional Pose Machines (CPM) [61] and Iterative Error Feedback (IEF) [58]. Since these methods assume that pedestrians have been detected *a priori*, we use different detectors to localize pedestrians: DPM, Faster R-CNN and varying degrees of jittered ground truth bounding boxes. We also test the ability of CPM and IEF to perform both detection and pose estimation simultaneously as a baseline *i.e.*, using the whole region as the input without localizing the pedestrian.

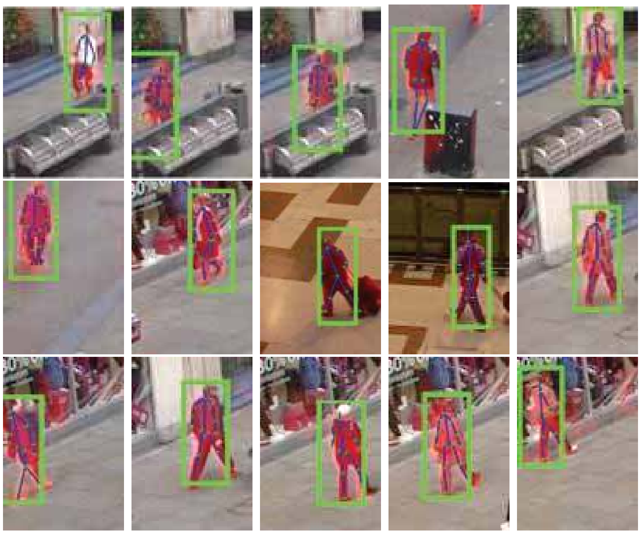


Fig. 15: Qualitative results of our approach predicting bounding box, body pose in terms of part locations (skeleton) and a (segmentation mask). The first row shows examples where the pedestrian is occluded.

4.3.2 Pedestrian Detection Evaluation

We compare our ScenePoseNet model to all baselines using the standard 50% overlap metric used for pedestrian detection. Although in theory we can learn a ScenePoseNet model for every location or region in the scene, pedestrians in the datasets tend to walk only in certain parts of the scene. For efficiency, we evaluate detection accuracy

Table 4: Mean average precision and mean IoU

Method	meanIoU	mAP
Ours	0.5502	0.768
SLSV [39]	0.4041	0.5201

on real pedestrians using only high traffic areas. Results are summarized in the precision recall (PR) curve in Fig. 14. The PR curves show that our approach has a significantly better recall rate due to our ability to learn accurate scene-and-region specific detectors.

Our approach, trained purely on synthetic data, outperforms generic state-of-the-art detectors that are trained on real data. This provides validation for our premise that explicitly making use of scene geometry, obstacles and camera setup can significantly help synthesis based techniques outperform models that are trained on real data. Finally we also compare the performance of our approach with SLSV, which also learns a scene scene-specific pedestrian detection model based on traditional HoG features. The comparison of mean average precision by 50% bounding box overlap and mean IoU at high traffic region in the Towncenter dataset are summarized in Table 4. The results show that ScenePoseNet exhibits better localization performance in comparison to SLSV even when both of the approaches leverage scene geometry. We believe that this is due in part to the ability of ScenePoseNet to learn both the features and regressor end-to-end.

4.3.3 Pose Estimation Evaluation

We compare our ScenePoseNet model to all baselines using the standard PCKh metric used for pose estimation. We evaluate pose estimation accuracy on real pedestrians. Results are summarized as a function of overlap threshold in Fig. 16a and Fig. 16b for the Towncenter and the PETS2006 datasets respectively. Our approach outperforms all the baselines on real pedestrian data from the two scenes without using any real data for training. By generating physically grounded and geometrically accurate renderings of pedestrians along with high-quality segmentation masks and noise free joint annotations, ScenePoseNet is able to bridge the gap between real and synthetic data. In Fig. 15, qualitative results are provided at different regions on both PETS2006 and Towncenter datasets.

Finally, we quantify the performance of just the pose estimator by presenting the baseline pose estimators with ground truth pedestrian detection and randomly jittered ground truth (to simulate a better detector). We also compare against a variant of ScenePoseNet (SPN-G), that learns only one *general* network for the entire scene and is not tuned to any specific region of the scene. Fig. 16c shows this comparison along with the SPN-G variant. The SPN-G variant that is trained for the entire dataset also outperforms the generic pose estimators.

4.3.4 Time Complexity

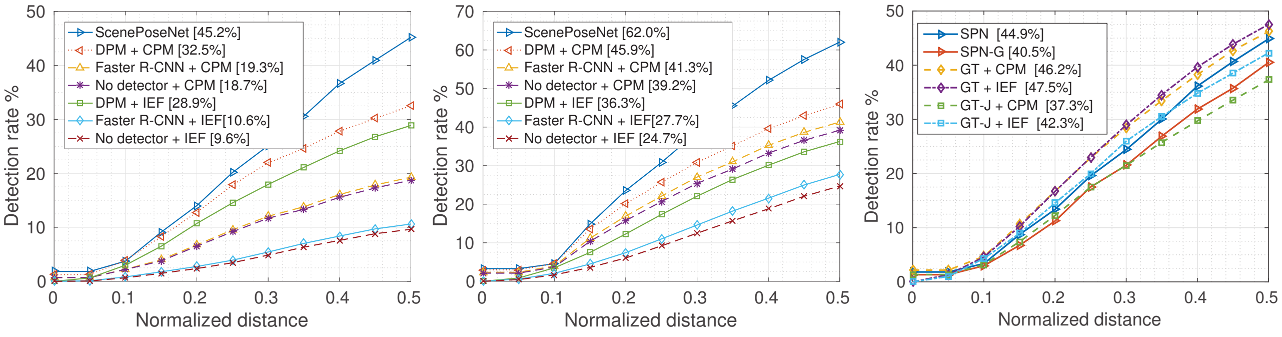
We compare the inference time complexity of Scene Pose Net and the baselines for detection, pose estimation and the combined task. The timing results are summarized in Fig. 18. We used code provided by the authors for the baselines and all timing measurements were performed on the same computational setup with an Intel i7-5390 processor with a single Titan-X GPU. The time for the joint task of detection and pose estimation depends on the respective detection and pose estimation baseline combinations. By coupling the tasks of human localization and pose estimation into a single network, ScenePoseNet is significantly, over 100%, faster than the fastest baseline combination, Faster-RCNN + CPM. ScenePoseNet processes each frame in 0.18sec while the baseline combination takes around 0.37sec for both detection and pose estimation.

4.3.5 Ablative Analysis

Effect of data: Here we study the effect of rendering data with prior knowledge and the amount of data being used. We perform the following comparisons (see Fig. 17a): (1) 50,000 training renders sampled from a prior distribution of pedestrian orientation and pose, (2) 50,000 training renders sampled from a uniform distribution of orientation and pose, and (3) 150,000 training renders sampled from a prior distribution of pedestrian orientation and pose. Leveraging prior knowledge on the likely orientation and pose of people in the scene allows us make effective use of the rendered data. Furthermore, we observed that using more than 50,000 training images does not improve the performance on real images, therefore we use 50,000 images to train our models, sampled from a prior distribution for the Towncenter and from a uniform distribution for the PETS2006 dataset.

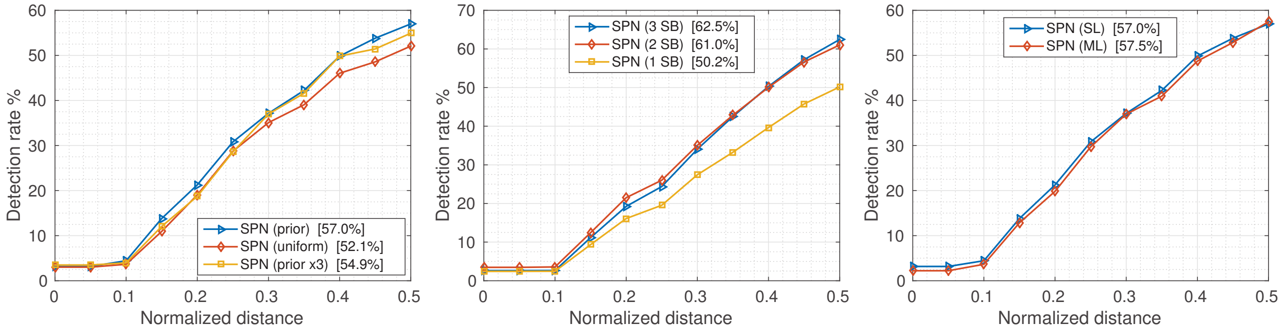
Effect of SB units: Here we study the effect of the number of stacked SB units (see Fig. 17b for the results). We observe that using two SB gives a significant boost over using a single SB unit. Adding one more SB unit does not seem to help with the synthetic data but provides a slight performance boost on real data.

Intermediate supervision and skip connections: We evaluate the effectiveness of using intermediate supervision, as suggested by some recent pose estimation



(a) Towncenter (b) PETS2006 (c) Ground Truth (Towncenter)

Fig. 16: Pose estimation performance on the (a) Towncenter and the (b) PETS2006 dataset against multiple baselines on real data. The number in the bracket corresponds to a PCKh threshold of 0.5. The baselines are combinations of state-of-the-art detection and pose estimation methods as well as pose estimation without pedestrian detection. (c) We also compare pose estimation performance on the Towncenter dataset when using the ground truth (GT) detections and their jittered (GT-J) versions as well as the ScenePoseNet generic– SPN-G, where we learn a single model for the entire scene.



(a) Data Prior (b) Number of Spatial Belief Units (c) Intermediate Supervision

Fig. 17: Pose estimation results on real data: (a) here we demonstrate the advantage of using a data prior for sampling pedestrian orientation and pose, (b) exploring the effect of number of spatial belief units, and (c) the effect of training our model with intermediate supervision i.e., optimizing a single loss function (SL) and multiple loss functions (ML).

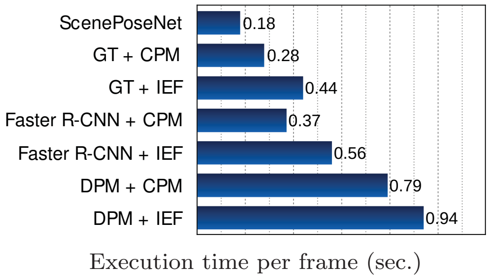


Fig. 18: Comparison of speed across different approaches for pose estimation and pedestrian detection.

approaches [61], on the performance of ScenePoseNet. We train and evaluate our network with intermediate supervision at the outputs of the spatial-belief units. Fig. 17c shows the comparison. The network using intermediate supervision only provides a small gain in performance. Therefore, we do not use intermediate supervision in our experiments. Finally, we note that the skip connections have proven critical in being able to train our network. Repeated attempts to train our network without the skip connections has resulted in convergence failure.

5 Conclusion

We have presented a purely synthetic approach to training scene-specific location-specific pedestrian detectors and pose estimators. We showed that by leveraging the parameters of the camera and known geometric layout of the scene, we are able to learn customized pedestrian models for every part of the scene. In particular, our proposed approach took into account the perspective projection of pedestrians on the image plane and also modeled pedestrian appearance under synthetic object occlusion.

Our proposed algorithm jointly learns hundreds of pedestrian models using an efficient alternating algorithm, which fine tunes each pedestrian detector while also enforcing spatial smoothness between models. Our experiments showed that our model outperforms several baseline approaches in terms of image plane localization and as well as localization in 3D.

For scaling to multi-task learning for further analysis, our algorithm generates a deep convolutional neural network trained on scene specific synthetic data. The rendering system generates physically grounded and geometrically plausible renders of synthetic humans that serve as training data for our scene-specific pedestrian detection and pose estimation model. Our experimental results suggest a surprising outcome that our approach can effectively generate a pedestrian detector and pose estimator just from a high level description of the scene. The models by our framework can serve as an alternative to using state-of-the-art off-the-shelf generic for pedestrian detection and pose estimation.

Synthesis-based training techniques are well suited for the current paradigm of data-hungry object detectors. Although we have focused primarily on the use of scene geometry for synthesis, it is only the first step in maximizing prior scene knowledge for synthesis. We have yet to explore the more high-level semantic interpretation of the scene which can be used to generate a wider range of human poses. For example, functional attributes of the scene provide strong priors on walking direction, probable pose and likely occlusion patterns which can be used to generate a wider range of synthetic images of people. We believe that advances in functional scene understanding and improvements in human rendering techniques will enable more powerful models using our detection-from-synthesis approach.

Acknowledgement Partially supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/ Interior Business Center (DOI/IBC) contract number D17PC00340. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation/herein. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.

References

1. A. Krizhevsky, I. Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1
2. J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1
3. A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 1
4. Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 1
5. Rodney A. Brooks. Symbolic reasoning among 3-d models and 2-d images. *Artificial Intelligence*, 17(13):285 – 348, 1981. 2
6. Michel Dhome, Ali Yassine, and Jean-Marc Lavest. Determination of the pose of an articulated object from a single perspective view. In *BMVC*, pages 1–10, 1993. 3
7. Alberto Broggi, Alessandra Fascioli, Paolo Grisleri, Thorsten Graf, and M Meinecke. Model-based validation approaches and matching techniques for automotive vision based pedestrian detection. In *CVPR Workshop*, pages 1–1. IEEE, 2005. 3
8. Kristen Grauman, Gregory Shakhnarovich, and Trevor Darrell. Inferring 3d structure with a statistical image-based shape model. In *ICCV*, pages 641–647. IEEE, 2003. 3
9. Ankur Agarwal and Bill Triggs. A local basis representation for estimating human pose from cluttered images. In *ACCV*. Springer, 2006. 3
10. Javier Marin, David Vázquez, David Gerónimo, and Antonio M López. Learning appearance in virtual scenarios for pedestrian detection. In *CVPR*, pages 137–144. IEEE, 2010. 3
11. Vassilis Athitsos, Haijing Wang, and Alexandra Stefan. A database-based framework for gesture recognition. *Personal and Ubiquitous Computing*, 14(6):511–526, 2010. 3
12. Michalis Potamias and Vassilis Athitsos. Nearest neighbor search methods for handshape recognition. In *Proceedings of the 1st international conference on Pervasive Technologies Related to Assistive Environments*, page 30. ACM, 2008. 3
13. Javier Romero, H Kjellstrom, and Danica Kragic. Hands in action: real-time 3d reconstruction of hands in interaction with objects. In *ICRA*, 2010. 3
14. Bojan Pepik, Michael Stark, Peter Gehler, and Bernt Schiele. Teaching 3d geometry to deformable part models. In *CVPR*, pages 3362–3369. IEEE, 2012. 3
15. Yair Movshovitz-Attias, Vishnu Naresh Boddeti, Zijun Wei, and Yaser Sheikh. 3d pose-by-detection of vehicles via discriminatively reduced ensembles of correlation filters. In *BMVC*, 2014. 3
16. Mohsen Hejrati and Deva Ramanan. Analysis by synthesis: 3d object recognition by object reconstruction. In *CVPR*, pages 2449–2456. IEEE, 2014. 3
17. Kevin Lai, Liefeng Bo, and Dieter Fox. Unsupervised feature learning for 3d scene labeling. In *ICRA*, 2012. 3
18. Scott Satkin, Jason Lin, and Martial Hebert. Data-driven scene understanding from 3d models. In *BMVC*, 2012. 3
19. Baochen Sun and Kate Saenko. From virtual to reality: Fast adaptation of virtual object detectors to real domains. In *BMVC*, 2014. 3

20. Geoffrey R Taylor, Andrew J Chosak, and Paul C Brewer. Ovvv: Using virtual worlds to design and evaluate surveillance systems. In *CVPR*, pages 1–8, 2007. 3
21. M. Aubry, D. Maturana, A. Efros, B. Russell, and J. Sivic. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In *CVPR*, 2014. 3
22. David Vazquez, A López, Javier Marin, Daniel Ponsa, and David Gerónimo. Virtual and real world adaptation for pedestrian detection. *PAMI*, 36(4):797–809, April 2014. 3, 10, 11, 13
23. L. Pishchulin, A. Jain, M. Andriluka, T. Thormählen, and B. Schiele. Articulated people detection and pose estimation: Reshaping the future. In *CVPR*, 2012. 3
24. German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio Lopez. The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016. 3
25. A Gaidon, Q Wang, Y Cabon, and E Vig. Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*, 2016. 3
26. Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *European Conference on Computer Vision (ECCV)*, volume 9906 of *LNCS*, pages 102–118. Springer International Publishing, 2016. 3
27. Pyry Matikainen, Rahul Sukthankar, and Martial Hebert. Classifier ensemble recommendation. In *ECCV Workshop*, pages 209–218. Springer, 2012. 3
28. Leonid Pishchulin, Arjun Jain, Christian Wojek, Mykhaylo Andriluka, T Thormahlen, and Bernt Schiele. Learning people detection models from few training samples. In *CVPR*, pages 1473–1480. IEEE, 2011. 3
29. Meng Wang and Xiaogang Wang. Automatic adaptation of a generic pedestrian detector to a specific traffic scene. In *CVPR*, pages 3401–3408. IEEE, 2011. 3
30. Meng Wang, Wei Li, and Xiaogang Wang. Transferring a generic pedestrian detector towards specific scenes. In *CVPR*, pages 3274–3281. IEEE, 2012. 3
31. Xiaogang Wang, Meng Wang, and Wei Li. Scene-specific pedestrian detection for static video surveillance. *PAMI*, 36(2):361–374, Feb 2014. 3
32. Jiaolong Xu, David Vázquez, Sebastian Ramos, Antonio M López, and Daniel Ponsa. Adapting a pedestrian detector by boosting lda exemplar classifiers. In *CVPR Workshop*, pages 688–693. IEEE, 2013. 3
33. Enver Sangineto. Statistical and spatial consensus collection for detector adaptation. In *ECCV*, pages 456–471. Springer, 2014. 3
34. Yang Yang, Guang Shu, and Mubarak Shah. Semi-supervised learning of feature hierarchies for object detection in a video. In *CVPR*, pages 1650–1657. IEEE, 2013. 3
35. Shiyu Huang and Deva Ramanan. Expecting the unexpected: Training detectors for unusual pedestrians with adversarial imposters. In *CVPR*, 2017. 3
36. H. Su, C.R. Qi, Y. Li, and L.J. Guibas. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In *ICCV*, 2015. 3
37. P. Fischer, A. Dosovitskiy, E. Ilg, P. Häusser, C. Hazırbaş, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *ICCV*, 2015. 3
38. J. Shotton, R.B. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, et al. Efficient human pose estimation from single depth images. *PAMI*, 2013. 3
39. H. Hattori, V.N. Boddeti, K.M. Kitani, and T. Kanade. Learning scene-specific pedestrian detectors without real data. In *CVPR*, 2015. 3, 13, 15
40. Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005. 4, 10, 11, 13
41. P. Dollár, Z. Tu, P. Perona, and S. Belongie. Integral channel features. In *BMVC*, 2009. 4
42. S. Zhang, R. Benenson, and B. Schiele. Filtered channel features for pedestrian detection. In *CVPR*, 2015. 4
43. Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *PAMI*, 34(4):743–761, 2012. 4, 11
44. Vishnu Naresh Boddeti, Takeo Kanade, and BVK Kumar. Correlation filters for object alignment. In *CVPR*, pages 2291–2298, 2013. 4, 6
45. Joao F Henriques, Joao Carreira, Rui Caseiro, and Jorge Batista. Beyond hard negative mining: Efficient detector learning via block-circulant decomposition. In *ICCV*, 2013. 4
46. Y. Tian, X. Wang P. Luo, and X. Tang. Deep learning strong parts for pedestrian detection. In *ICCV*, 2015. 4
47. Z. Cai, M. Saberian, and N. Vasconcelos. Learning complexity-aware cascades for deep pedestrian detection. In *ICCV*, 2015. 4
48. R.B. Girshick, P.F. Felzenszwalb, and D.A. Mcallester. Object detection with grammar models. In *NIPS*, 2011. 4
49. W. Ouyang and X. Wang. Joint deep learning for pedestrian detection. In *ICCV*, 2013. 4
50. R. Girshick. Fast r-cnn. In *ICCV*, 2015. 4
51. S. Ren, K. He, R.B. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 4, 14
52. W. Liu, D. Anguelov, D. Erhan, and C. Szegedy. SSD: Single shot multibox detector. In *ECCV*, 2016. 4
53. P.F. Felzenszwalb and D.P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 2005. 4
54. Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *PAMI*, 2013. 4, 13
55. L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Strong appearance and expressive spatial models for human pose estimation. In *ICCV*, 2013. 4
56. W. Yang, W. Ouyang, H. Li, and X. Wang. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In *CVPR*, 2016. 4
57. A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 2014. 4
58. J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. In *CVPR*, 2016. 4, 14
59. Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from Synthetic Humans. In *CVPR*, 2017. 4
60. V. Ramakrishna, D. Munoz, M. Hebert, A.J. Bagnell, and Y. Sheikh. Pose machines: Articulated pose estimation via inference machines. In *ECCV*, 2014. 4
61. S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016. 4, 9, 14, 16
62. Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. *arXiv preprint arXiv:1603.06937*, 2016. 4

63. Grégory Rogez and Cordelia Schmid. Mocap-guided data augmentation for 3d pose estimation in the wild. In *NIPS*, 2016. 4
64. Andreas Ess, Bastian Leibe, and Luc Van Gool. Depth and appearance for mobile scene analysis. In *ICCV*, pages 1–8, 2007. 4
65. Markus Enzweiler and Dariu M Gavrila. Monocular pedestrian detection: Survey and experiments. *PAMI*, 31(12):2179–2195, 2009. 4
66. Christian Wojek, Stefan Walk, and Bernt Schiele. Multi-cue onboard pedestrian detection. In *CVPR*, pages 794–801, 2009. 4
67. Biswajit Bose and Eric Grimson. Improving object classification in far-field video. In *CVPR, 2004.*, volume 2, pages II–II. IEEE, 2004. 4
68. Peter M Roth, Sabine Sternig, Helmut Grabner, and Horst Bischof. Classifier grids for robust adaptive object detection. In *CVPR*, pages 2727–2734. IEEE, 2009. 4
69. Severin Stalder, Helmut Grabner, and LV Gool. Exploring context to learn scene specific object detectors. In *Proc. PETS*, 2009. 4
70. Severin Stalder, Helmut Grabner, and Luc Van Gool. Cascaded confidence filtering for improved tracking-by-detection. In *ECCV, 2010.*, pages 369–382. Springer, 2010. 4
71. Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011. 7
72. K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. *arXiv preprint arXiv:1603.05027*, 2016. 8, 9
73. Ben Benfold and Ian Reid. Stable multi-target tracking in real-time surveillance video. In *CVPR*, pages 3457–3464, 2011. 10
74. David Thirde, Longzhen Li, and F Ferryman. Overview of the PETS2006 challenge. In *Proc. 9th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS 2006)*, pages 47–50, 2006. 10
75. Koosuke Hattori, Hironori Hattori, Yuji Ono, Katsuaki Nishino, Masaya Itoh, Vishnu Naresh Boddeti, and Takeo Kanade. Carnegie Mellon University Surveillance Research Dataset (CMUSRD). Technical report, Carnegie Mellon University, Nov. 2014. `http://www.consortium.ri.cmu.edu/projSRD.php`. 10
76. P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9):1627–1645, 2010. 10, 11, 13
77. R. B. Girshick, P. F. Felzenszwalb, and D. McAllester. Discriminatively trained deformable part models, release 5. `http://people.cs.uchicago.edu/~rbg/latent-release5/`, 0000. 10
78. Derek Hoiem, Alexei A Efros, and Martial Hebert. Putting objects in perspective. *IJCV*, 80(1):3–15, 2008. 10, 11, 13