Priming Deep Pedestrian Detection with Geometric Context

Anonymous

***Abstract*— We investigate the role of geometric context in deep neural networks to establish better pedestrian detectors that are more robust to occlusions. Notwithstanding their demonstrated successes in the wild, deep object detectors underperform in crowded scenes with high intra-category occlusions. One brute-force solution is to collect a large number of labeled training samples under occlusion, but the combinatorial increase in the labeling effort makes it an unaffordable solution. We argue that a promising and complementary direction to solve this problem is to bring geometric context to modulate feature learning in a DNN. We identify that an effective way to leverage geometric context is to induce it in two steps - through early fusion, by guiding region proposal generation to focus on occluded regions, and through late fusion, by penalizing misalignments of bounding boxes in both 2D and 3D. Our experiments on multiple state-of-the-art DNN detectors and several detection benchmarks clearly demonstrates that our proposed method outperforms strong baselines by an average of 5%.**

I. Introduction

Humans have an innate understanding of geometry [1]. Concepts such as occlusion, perspective and groundedness help us to represent objects, encode their locations, and mentally manipulate them in the physical world [2]. Early research in object recognition was also founded on these principles when formal geometric models were used as informative priors to modulate statistical learning.

Of late, object detectors have been increasingly cast as regression on deep convolutional image features. Neural networks such as R-FCN [3] and Faster RCNN [4] are trained on large datasets, making them robust to most perceptual variations. Nevertheless, they often fail in crowded scenes with large number of instances of the same category, as shown in Figure 1. On the left, an R-FCN model performs reasonably well in detecting pedestrians in the scene. However, if we zoom into a crowded corner, we notice that pedestrians are often missed completely or manifest as merged detections, with multiple instances detected as a single blob. Delineating these partially visible instances could increase the overall detection accuracy and improve downstream tasks such as people tracking and activity recognition.

One of the key reasons for sub-optimal performance in crowded scenes is that the object detectors are trained on the statistics of appearance features alone. Typically, these detectors constitute a multi-stage pipeline through which images are first parsed into salient regions and then refined into bounding boxes, all of which is confined to the 2D image (pixel) space. The assumption is that *i.i.d.* sampling from a large training data would eventually cover all possible variations in the viewspace, resulting in an occlusion resilient model. These convolutional features do not account for scene regularities such as expected object size at a certain location, or that most objects are grounded.

Objects in the physical world are a lot more persistent than what appears on the retina or a camera sensor. Yet, humans are able to see a stable percept by correlating visual inputs to the underlying scene geometry. Sizes of familiar objects are correctly resolved based on knowledge of scene perspective. Partial occlusion is compensated by the process of *amodal completion*, i.e., by filling in information behind occluding surfaces. Inspired by human vision, in this work we posit that 2D image projection of neighboring objects in 3D is a consequence of the supporting scene geometry, i.e., the object appearance and occlusion patterns depends on the camera viewpoint with respect to the ground plane on which the objects rest. Following this physical view of image formation, we develop a detection model for pedestrians that factors in the so-called geometric context of the scene to account for complex feature patterns during model training.

Our model consists of a proposal-based deep neural network that is modulated by the geometric context of the scene. The *geometric context* acts as a prior at two levels of learning - (a) early fusion, through *occluder-centric proposals* to model inter-person occlusion during region proposal generation (RPN) and (b) late fusion, through *geometric loss* that models object position and scale consistency during per region loss computation. Our main contributions are:

- **A principled approach to incorporate geometric context into DNN models**. We factor in geometric context of the scene through early fusion, during region proposal generation and through late fusion, through loss computation, while maintaining the feedforward topology of the underlying object detector.
- **Occluder-centric proposals under variable viewpoint**. We augment dense, viewpoint-agnostic region proposals with sparse, occluder-centric proposals during region proposal generation, to focus on scene-specific hard regions, while learning to detect under diverse set occlusion patterns.
- **Geometric loss for bounding box regression.** We introduce geometric loss for bounding box regression, where we use the scene layout as a projection space to impose relational and position consistency of proposals on the ground plane.

To the best of our knowledge, this is the first approach that brings these concepts together systematically for pedestrian detection. Our experiments on state-of-the-art DNN detectors and detection benchmarks demonstrates that our proposed method outperforms strong baselines by an average of 5%.

Preprint submitted to 2019 International Conference on Robotics and Automation (ICRA). Received September 15, 2018.

Fig. 1: Standard detectors under-perform in crowded scenes. We factor in scene geometry to recover occluded instances. *Right*: The occlusion pattern of nearby objects in the image depends on the camera viewpoint *w.r.t.* the ground plane in 3D.

II. Related Work

In the last decade, several works have attempted to incorporate geometric context to improve statistical object detectors [5], [6], [7], [8], [9], [10]. Hoiem *et al.* [5] proposed a graphical model that placed local object detection in the overall 3D scene by modeling the interdependence of object sizes, surface orientations and camera viewpoint. This notion was extended in [6] to include additional pose constraints on object configurations within a feedback. Hedau *et al.* [8] proposed a cubic room representation for indoor scenes, showing that layout estimation improves object detection. These models improve the overall accuracy, but are usually staged as sequential or iterative components that couple with an independent object detector to refine object hypotheses.

At present, most accurate object detectors are predominantly single stage [11], [12] or multi-stage convolutional neural networks [4], [3] that behave as highly parallelized sliding window classifiers. Context inclusion as structured prediction becomes harder to implement without breaking the end-to-end formulation of the fixed CNN architecture. Some preliminary work have explored workarounds. In [13], object masks from current detection are fed as contextual input during subsequent iterations. To induce scale context, Foveanet [14] generates a perspective heatmap and fuses inference from a coarse and a fine subnetwork. These methods are limited to adapting for scale under perspective distortions.

One inventive way of inducing geometric context in deep learning is through reprojection loss, that quantifies how closely a 3D point estimate matches its true projection in 2D image space. It was introduced for single view depth estimation in [15], and since then has been extensively used for other self-supervised learning such as inferring egomotion and depth [16], [17], optical flow [18] and localization [19]. Typically, these methods presume a given or an inferred common reference frame. Inspired by these approaches, in this work, we introduce reprojection loss into object detection. We establish the ground plane as the common frame of reference and measure position consistency in 3D.

Occlusion reasoning in humans is powered by an innate ability to perform "amodal completion" [21], i.e., functional completion of partially occluded object shapes. In artificial vision, Kar et al. [22] have considered amodal box completion to find veridical sizes of objects. Hsiao and Hebert [23] developed an occlusion model by reasoning about objects as probabilistic 3D blocks. Inspired by Reverse Hierarchy Theory [24], Huang and Murphy [25] factorize an image into a bottom-up foreground object followed by top-down generation of the hidden parts. All these methods treat amodal completion purely as a learning problem. In contrast, we show that amodal completion is essentially a geometric phenomenon that relies on the underlying scene cues.

Several works, including Hoiem *et al.* use pedestrian detection as a case study for analyzing effects of geometric modulation on statistical detection. Pedestrian detection in crowds remains an active field with research and has spawned learning techniques including hard example mining [26], recurrent non-maxima suppression [27], synthesized datasets [28], and learning in multi-view, fully calibrated setting [29]. Our method is complementary to these approaches. We show that by utilizing the underlying scene geometry to modulate feature activation, we can learn discriminative features that improve the baseline pedestrian detector.

III. Model Overview

Our goal is to build a statistical pedestrian detector that is aware of the underlying scene geometry, and is robust to intra-category occlusions. To realize it, we utilize *geometric context* of the scene, which refers to the camera pose *w.r.t.* to the ground plane. We hypothesize that inducing geometric context into feature learning will better model occluded parts and improve recall. To achieve this, we build upon the well-known architectures of two-stage object detection, namely Faster RCNN [26] and Region-based Fully Convolutional Network (R-FCN) [3], to which we induce geometric context during feature activation and loss computation.

Our pipeline is illustrated in Figure 2. If camera pose is unavailable, we perform a one-time, approximate camera pose estimation as a pre-processing step. Following that, we build upon the architecture of a typical two-stage object detector. During anchor generation, we augment dense, pixel based anchors with 2D projections of *sparse, occluder-centric anchors in 3D* (step (a) in Figure 2). Next, the anchors are

Preprint submitted to 2019 International Conference on Robotics and Automation (ICRA). Received September 15, 2018.

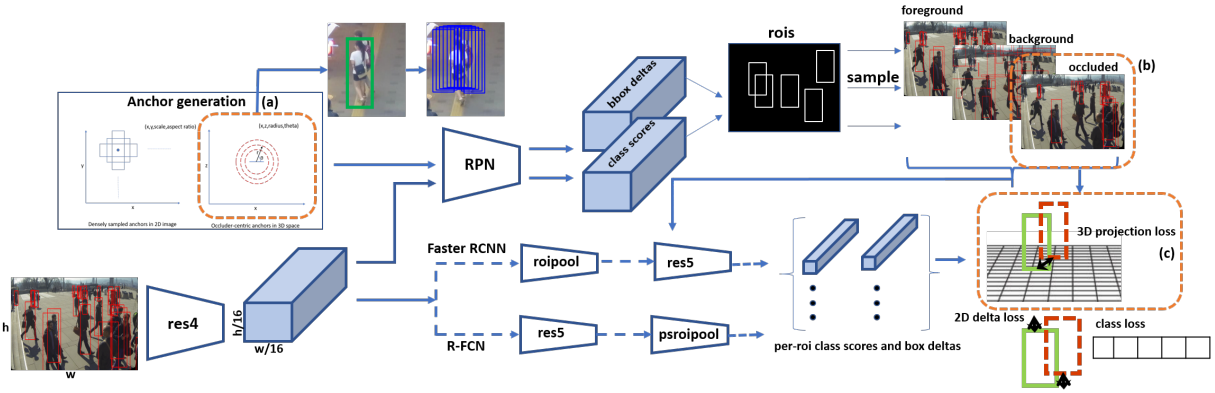


Fig. 2: Our object detection pipeline. We incorporate geometric cues into a standard object detector by early fusion, through - (a) occluder-centric anchors and (b) location sensitive RoIs, and late fusion through (c) 3D projection loss.

passed through the RPN stage, where they are filtered and labeled. In addition to foreground and background RoIs, we also *subsample a set of occluded anchors* based on their positions in 3D (step (b)). Finally, we jointly minimize the multi-task objective function by minimizing the location errors of RoIs on the ground plane. This is achieved by reprojecting the RoIs and ground-truth back onto the 3D space (step c).

We describe our geometric pre-processing step in Section IV. The design of occluder-centric anchors and proposal sampling is described in Section V. The geometric loss function is described in Section VI. Finally, experiments are described in Section VII, followed by conclusions in Section VIII.

IV. Estimating Geometric Context

Our work focuses on scenarios where a static camera captures wide-angle views of crowded scenes. In such scenarios, performing a simple, one-time camera pose estimation can provide rich spatial cues about the objects in space. In particular, knowing the camera pose *w.r.t.* the ground plane allows us to perform depth ordering of objects and transform points between image plane and ground plane, upto scale. Below, we briefly describe a geometric algorithm to compute camera extrinsics from scene cues.

We parameterize the camera projection matrix $P = [R, t]$ based on its angle of view and the height above ground plane. We solve these parameters using well known concepts from single view geometry [30]. We assume a pinhole camera model observing a Manhattan world, and known camera intrinsics K. For the camera angle, we annotate a few line segments to compute the vanishing point in z-direction v_z. Assuming zero roll, the horizon line is the corresponding image row v_0. The rotation vector is $\hat{r_3} = K^{-1}v_z/||K^{-1}v_z||$, which can be matched to the rotation matrix to recover the yaw = $tan^{-1}(\hat{r_{31}}/\hat{r_{33}})$ and pitch = $cos^{-1}(\hat{r_{32}})$ angles. For the camera height, we use the method introduced in Hoiem *et al.* [5]. Specifically, let y_c denote the camera height in world coordinates. Given the image height h_i, bottom position v_i, and world (physical) height y_i of a known object, the camera height y_c is computed using this formulation $y_c = y_i(v_i - v_0)/h_i$. We average the camera height estimates over multiple object instances in the scene to get a robust estimate.

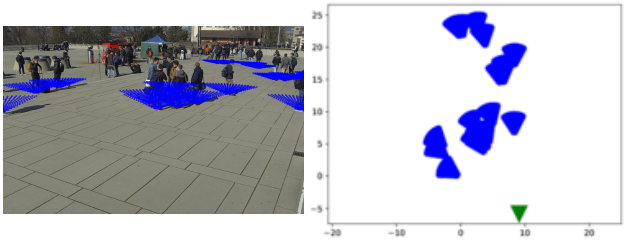


Fig. 3: Visualization of occlusion cones of pedestrians in a scene. *Left:* 2D projections in image space, *Right:* Bird's eye (top) view of the occlusion cones. Green triangle shows the camera position.

V. Geometry-guided Proposal Sampling

Region proposals in two-stage object detectors mimic sliding window search across all scales and aspect ratios. They are drawn from fixed set of anchors that are constant across all images. During RPN stage, anchors are classified into foreground and background classes based on their degree of overlap with ground-truth boxes. In addition to the foreground and background classes, we introduce a third category of anchors that adaptively focus attention towards occluded instances. By doing this, we explicitly prioritize location sensitive occluded RoIs during training.

We propose *occluder-centric anchors* that are identified by their positions relative to foreground occluders and scene-specific layout. In particular, occlusion occurs when multiple objects appear along the *line of sight* (*LoS*) of the camera at different depths. A cone shaped space behind the occluder encloses the region of occlusion in 3D. An object occupying this region manifests as partially visible in 2D [31]. Figure 3 shows examples of occlusion cones of pedestrians projected on the ground plane. In our proposed model, occluder-centric anchors are dense anchors that appear within occlusion cones. The steps to identify these are detailed below.

We first start by projecting 2D bounding boxes onto the ground plane using the known or inferred camera matrix.

Preprint submitted to 2019 International Conference on Robotics and Automation (ICRA). Received September 15, 2018.

Let (x_o, z_o) be a box's touch point on the ground. The *LoS* is the ray connecting the camera center to this point. Let the angle of this *LoS* be $\theta_{los,o}$. Then the occlusion cone corresponding to this box can be parameterized as (x_o, z_o, r, θ), where r denotes the radial distance from o and $\theta = \{\theta_{los,o} - \theta_\Delta, ..., \theta_{los,o} + \theta_\Delta\}$ denotes the angular separation from the *LoS*.

Next, at each parameterized location (x_o, z_o, r, θ), we generate hypothetical RoIs corresponding to expected human dimensions. Specifically, given (r, θ), the 3D positions of top left and bottom right positions in 3D are $(x_o - w/2, h, z_o)$ and $(x_o + w/2, 0, z_o)$, where (h, w) are expected height and width of humans (h=5.6 feet, w=1.7 feet, approx.). These 3D boxes are re-projected back onto image plane in form of 2D RoIs. The anchors that overlap these RoIs are indexed by corresponding (r, θ) values. Finally, a subset of these anchors are uniformly sampled from the (r, θ) distribution.

After proposal annotation stage, let $\{b_{fg}, b_{occ}, b_{bg}\}$ be the random subsets of foreground, occluded and background anchors. The combined loss function at the region classifier stage is expressed as follows.

$$\mathcal{L}_\mathcal{F} = \lambda_{fg}\mathcal{L}_\sigma^{l1}(\mathcal{F}(b_{fg}^{(c)}), C) + \lambda_{occ}\mathcal{L}_\sigma^{l1}(\mathcal{F}(b_{occ}^{(c)}), C) + \lambda_{bg}\mathcal{L}_\sigma(\mathcal{F}(b_{bg})), \quad (1)$$

where λ is the appropriate scaling function per proposal set, $\{\sigma, l1\}$ denotes the combination of softmax loss for class probability and smooth-L1 loss for bounding box regression between a proposal and the corresponding ground-truth box, C. Background proposals do not carry box regression penalty.

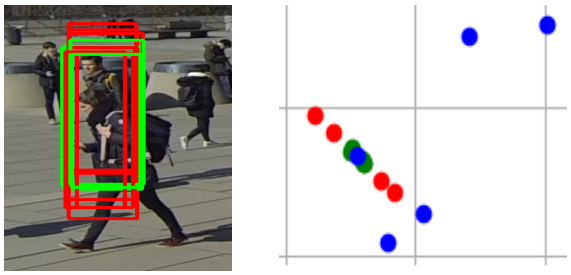


Fig. 4: Red boxes have IoU>0.7 with groundtruth. Green boxes have IoU>0.7 and Euclidean distance<$50cm$. *Right* shows the top view, with groundtruths as blue dots. Green boxes fit better, both in image and on ground, while the red boxes that are equidistant from multiple ground-truths lead to incorrect localization.

VI. 3D Projection Loss

In Equation 1, a multi-task loss $\mathcal{L}_\sigma^{l1}$ includes a classification loss that penalizes object category mismatch, and a regression loss that penalizes offsets of an anchor box to a nearby ground-truth in pixels. In addition, we design a geometric loss that encodes the position consistency of anchors on the ground. The particulars of our method is described below.

In Faster-RCNN, the regression target is parameterized into four dimensions through a convolutional subnetwork $bbox\_reg$ as $t_x = (x_g - x_a)/w_a, t_y = (y_g - y_a)/h_a, t_w = log(w_g/w_a), t_h = log(h_g/h_a)$, where $(.)_g$ and $(.)_a$ indicate ground-truth and anchor, respectively, and (x, y) are the box mid-points and (w, h) are the corresponding width and height. The technique in R-FCN is similar, except that a position sensitive maps are used as intermediate representations. In both approaches, anchor boxes are matched to ground-truth boxes based on Intersection over Union (IoU) overlap. The $bbox\_reg$ subnet is expected to learn a small offset that "corrects" the anchor position and aligns them perfectly. However, good anchor candidates might be farther from the object than can be learnt through the corrective offsets, as shown in Figure 4. The red boxes have significant overlap with the selected ground-truth, but are equidistant to neighboring ground-truths. This leads to coarse localization as multiple ground-truth boxes compete for the same anchor set.

Given the camera pose w.r.t. the ground plane, we introduce a projection loss function that attempts to resolve these issues by expanding the regression criterion. To achieve this, we add two additional dimensions to the regression target that predict the normalized distance between anchor and object in x_w and z_w, i.e., $t_x^w = (tpx_g - tpx_a)/g_x$ and $t_z^w = (tpz_g - tpz_a)/g_z$, where $tp_{(.)}$ denotes the touch points of bounding boxes projected in the x-z dimensions of the ground plane. In Figure 4, the green boxes are RoIs that have $IoU > 0.7$, and are also within $0.5m$ from the ground-truth box. As evident, these RoIs fit the object better.

In order to establish targets for regression, we need to project 2D box coordinates (x, y) to locations (x_w, y_w, z_w) on the ground plane. While exact recovery of 3D from 2D points is ill-posed, we make the solution tractable by making two reasonable assumptions - (a) pedestrians rest on ground plane, i.e., $y_w = 0$, and (b) the direction of ground plane is parallel to an upright person, i.e., $\hat{\mathcal{G}} = (0, 1, 0)$. Based on these constraints, the position of an RoI on the ground plane (denoted by $\mathcal{G}(b)_{y_w=0}$) is computed algebraically as follows - (a) find the camera center, given by the null space of the camera matrix P = $null(P)$. (b) The ray emanating from the camera center to the object is given by $\vec{R} = inv(P) * X_{im}$, where X_{im} is the bottom coordinate of an image box, and finally (c) the intersection between the projected ray $\vec{R}$ and the ground unit vector $\hat{\mathcal{G}}$ is computed algebraically to give the corresponding touch point X_g on the ground.

Given 3D position of boxes, we can evaluate the 2D + 3D regression loss. Let the set of foreground and occluded anchors, $\{b_{fg}, b_{occ}\}$ as computed in Section V, and the corresponding ground-truth boxes be C. Then the offsets are penalized through a smooth L1 function. The predicted feature map $\mathcal{F}$ are expanded to accommodate regression terms along the two dimensions, normalized by the width and height of the ground plane, respectively.

$$\mathcal{L}_{2D+3D} = \sum_{b \in n_R} \mathcal{L}_{l1}(\mathcal{F}_{2D}(b), t_{b,2D}^g) + \mathcal{L}_{l1}(\mathcal{F}_{3D}(b), t_{b,3D}^g), \quad (2)$$
$$\text{where,} \quad t_{b,3D}^{g,w} = \{[\mathcal{G}(b_{gt}^w)]_{y_w=0} - [\mathcal{G}(b_a^w)]_{y_w=0}\}/W,$$

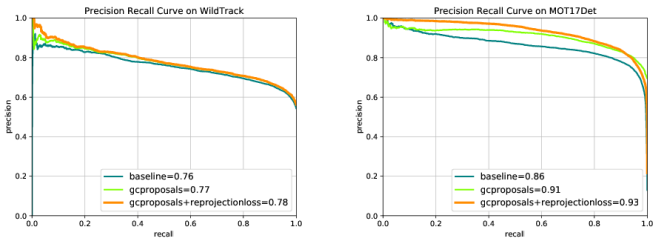


Fig. 5: Precision-recall curves of geometric-RFCN model on ETH Wildtrack dataset (*left*), and on MOT17 Detection dataset (*right*).

VII. Experiments

This section is organized as follows: we first introduce the datasets and the experiment settings; then we evaluate our proposed algorithm vis-a-vis the baselines; and finally we perform controlled experiments to study the impact of individual design choices.

A. Experiment design

We evaluate our algorithm on two pedestrian datasets - ETH WildTrack [32] and MOT 2017 Detection (MOT) [33]. WildTrack dataset consists of seven sequences of pedestrians from overlapping viewpoints with variable camera heights. The cameras were jointly calibrated and the corresponding camera intrinsics and extrinsics are available. To test our method on ambient scenes, we evaluate on the MOT dataset, which consists of seven sequences of various indoor and outdoor pedestrian traffic. We perform scene calibration using single view geometry (explained in Section IV) and use the inferred geometric parameters in our experiments. Both datasets capture video frames of crowded, dynamic scenes of moving pedestrians (~17 pedestrians per frame), using static cameras mounted at a variety of vantage points (e.g., surveillance view, below eye-level view, etc.) and provide full body annotations on all groundtruth instances. For evaluation, we partition each video temporally such that the first 60% of the frames is used for training and validation, while the remaining 40% is sequestered for testing.

We evaluate our approach on two well-studied architectures - Region based Fully Convolutional Network (R-FCN) [3] and Faster-RCNN [26], and build our algorithm on top of these pipelines.

mAP	cam1	cam2	cam3	cam4	cam5	cam6	cam7	avg
rfcn [3]	78.0	73.9	77.6	60.9	82.0	74.2	72.5	75.8
g-rfcn	**79.5**	**76.0**	**77.8**	**66.1**	**84.4**	**74.9**	**74.0**	**78.2**
frcnn [26]	78.4	**81.2**	82.9	61.1	84.0	75.5	84.7	78.8
g-frcnn	**80.4**	80.6	**83.4**	**69.7**	**85.2**	**77.9**	**84.7**	**80.1**

TABLE I: mAP per camera view on Wildtrack dataset. Prefix g- refers to our geometry primed model.

Implementation Details: All the training experiments share the same protocol. We use Resnet-101 as the backbone. The models are pre-trained on COCO dataset, and finetuned by freezing all layers uptil 4^{th} resnet block, and re-training all the layers thereafter. We train using SGD with momentum of 0.9, a weight decay of 5e-4, an initial learning rate of 5e-5 and with a batch size of 1. Each training routine is allowed to run for 20 epochs. We do not perform hyperparameter tuning or data augmentation. We use Intersection over Union (IoU) of 0.7 as threshold for positive detection and mean average precision (mAP) as evaluation metric.

mAP	mot2	mot4	mot5	mot9	mot10	mot11	mot13	avg
rfcn [3]	89.3	86.7	90.4	92.8	89.7	**96.8**	77.7	86.2
g-rfcn	**90.9**	**91.3**	**91.8**	**94.3**	**91.3**	96.3	**83.4**	**93.0**
frcnn [26]	89.7	89.5	92.0	94.9	88.1	96.9	78.3	90.2
g-frcnn	**90.4**	**92.7**	**93.4**	**96.1**	**89.2**	**97.2**	**81.5**	**92.4**

TABLE II: mAP per camera view on MOT dataset.

We set a budget of 256 RoIs at the RPN stage. In a baseline models, the proportion of foreground RoIs is set to 25%, and remaining 75% are background. In our proposed model, occluded-RoIs account for 12.5% of total RoIs, making the ratio of foreground, occluded and background RoIs as 25:12.5:62.5. Occluded RoIs are sampled from (r, θ) distribution using uniform sampling. The occluded anchors are densely sampled by varying the distance to the occluder (r) and orientation along the line of sight (θ). Depending on the camera height, the radius ranges from 1-10 feet for overhead camera view, and 1-50 feet for horizontal and below eye-level views, at increments of 1 foot. The angle of deviation from *line of sight* ranges between 0 to 60 degrees, at increments of 5 degrees. In order to project the anchors back onto the image plane, we assume an average human height of 5.5 feet for MOT and 5.9 feet for Wildtrack (provided in the dataset).

B. Comparison with the baselines

Overall accuracy: Using our proposed model, we achieve a higher mAP across both datasets and detection architectures. The baseline R-FCN achieves 75.24% on Wildtrack and 86.20% on MOT (See Tables I and II, and Figure 5). We improve these scores to 93.02% on MOT and 78.16% on Wildtrack, an improvement of 8% and 3%, respectively. Overall the performance of Faster RCNN is better than R-FCN, as have been observed before [3]. We outperform the baseline Faster RCNN model by 2% on an average. Tables I and II show the mAPs per camera view. We consistently outperform the baseline on all camera views. One interesting point to note is that mAP gains are higher for overhead views than for horizontal (eye-level) views. This is discussed in more detail in Section C.

Comparison on occluded samples: According to our hypothesis, occluded samples should be better detected by our geometry primed method. The results are shown in Table III. Indeed, we observe a 17 percentage points (pp) gain on Wildtrack and a 9 pp gain on the MOT dataset. The gains in the occluded segments are significant higher than the gains across entire data. This indicates that the occluded instances are better delineated using our method. A fine-grained analysis of the gains as a function of geometric parameters is studied

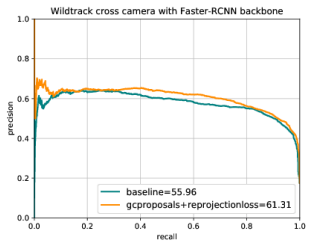
Preprint submitted to 2019 International Conference on Robotics and Automation (ICRA). Received September 15, 2018.

in Section C.

	cam1	cam2	cam3	cam4	cam5	cam6	cam7
baseline [3]	35.7	26.2	47.4	10.3	34.5	19.7	49.1
ours	**39.4**	**29.0**	**52.0**	**16.8**	**38.0**	**20.5**	**55.7**
	mot2	mot4	mot5	mot9	mot10	mot11	mot13
baseline [3]	42.6	41.9	50.4	26.9	23.8	50.3	4.7
ours	**45.0**	**52.6**	**51.4**	**27.8**	**25.1**	**50.7**	**5.8**

TABLE III: mAP over occluded instances. We improve by 17 pp ans 9 pp on Wildtrack and MOT datasets, respectively.

Cross camera performance: We evaluate cross-camera performance on the Wildtrack dataser to test if a superior model trained on a subset of views might also perform better on the unobserved views capturing the same underlying scene. We train on camera views C1-C4, and tested on C5-C7. As shown in Table IV, we improve the accuracy by 2% on the unobserved views over baseline. We also evaluate the accuracy on ground plane by measuring the average Euclidean distance between predicted locations and ground-truth in cm. We achieve a decrease of 15.3% in localization error compared to the baseline, as shown in Table IV.



mAP	C5	C6	C7
frcnn	62.9	47.5	60.8
g-frcnn	66.4	48.5	62.7
dist (cm)	C5	C6	C7
frcnn	148.4	883.4	60.7
g-frcnn	139.2	874.2	56.4

TABLE IV: Cross-camera performance on Wildtrack dataset. *Right below*: The prediction error on ground plane is shown as average Euclidean distance (lower is better).

C. Controlled experiments

Effect of geometric parameters: We analyze the mAP gains by camera elevation, inter spatial distance and radial angles between occluded objects on ground plane (Figure 6). In (a), we observe that camera elevation (x-axis) has a significant influence on the mAP gain. The higher the camera, the better our model is able to represent occluded patterns. At eye level, a person might get completely occluded, thus making it harder to model the relationship using geometric constraints. For example, views MOT11, WT4 and WT6 that are captured at eye-level show lower mAP gains. This argument also justifies why most surveillance cameras prefer overhead views for better scene capture. On the right in Figure 6 we plot the gain *w.r.t.* angle of deviation from the *line of sight*, where $los = 0^o$ is directly behind the occluder. Low-level and overhead camera views show similar patterns - we achieve higher gains at slight angular separation of ~15^o. Intuitively, these samples are challenging to learn without explicit attention provided by the geometric context. In contrast, at eye-level we see gains at larger radial separation, i.e., for side-by-side configuration.

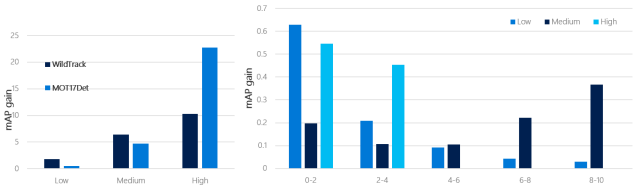


Fig. 6: mAP gains under occlusion *w.r.t.* camera elevation and radial distance.

mAP	pretrained coco	pretrained occluded RoIs	finetune	finetune+ [34]	train+	train + 3d
MOT	82.23	**84.38**	86.23	90.07	**91.29**	**93.02**
WT	48.19	**68.46**	75.87	75.95	**77.29**	**78.16**

TABLE V: Ablation Study

Ablation study: We evaluate the impact of individual design choices and modules in our network (Table V). In the first experiment, we consider using occluder-centric proposals during inference alone (columns 1 and 2). Specifically, we run two rounds of inference with pre-trained COCO model. In the second round, we use the top K detections from the first round and generate occluder-centric proposals centered at these detections for the second round. Remarkably, this proves to be a very effective strategy. We achieve a huge 41 point gain, from 48% in first round to 68% with augmented proposals (column 2). The gains are smaller for MOT, possibly because pedestrians appear in a variety of poses that violate the occlusion patterns. Next, we evaluate our model by training with and without Online Hard Negative Mining [34] (columns 3 and 4). OHEM retroactively selects a subset of RoIs that have the highest loss and backpropagates based on the selected examples. In contrast, we proactively select hard (occluded) ROIs. Our proposed models perform better than OHEM. Finally, we evaluate the impact of occluder-centric proposals alone (early fusion of geometry) and the combined effect of occluder-centric proposals along with 3D geometric loss (early + late fusion) (columns 5 and 6). Each module individually improves the mAP, but the best performance is achieved by using both the modules together (column 6).

VIII. Conclusion

In this paper, we incorporate geometric context into DNN based object detectors using approximate camera pose w.r.t. ground plane. Our method's novelty lies in a new anchor generation technique to sample occluder-centric proposals. We also propose to include geometric loss during bounding box regression. Ultimately, our algorithm leads to higher accuracy in monocular pedestrian detection. In the future, we will build upon this technique to investigate multi-view tracking and people re-identification.

Preprint submitted to 2019 International Conference on Robotics and Automation (ICRA). Received September 15, 2018.

REFERENCES

[1] S. Dehaene, V. Izard, P. Pica, and E. Spelke. Core knowledge of geometry in an amazonian indigene group. *Science*, 2006.

[2] N. S. Newcombe and J. Huttenlocher. *Making space: The development of spatial representation and reasoning*. MIT Press, 2000.

[3] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. *NIPS*, 2016.

[4] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[5] D. Hoiem, A.A. Efros, and M. Hebert. Putting objects in perspective. 80(1), October 2008.

[6] M. Sun, S.Y. Bao, and S. Savarese. Object detection using geometrical context feedback. 100(1), October 2012.

[7] D. Lee, A. Gupta, M. Hebert, and T. Kanade. Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. In *NIPS*, 2010.

[8] V. Hedau, D. Hoiem, and D. Forsyth. Recovering the spatial layout of cluttered rooms. In *ICCV*, 2009.

[9] J. Pan and T. Kanade. Coherent object detection with 3d geometric context from a single image. In *ICCV*, 2013.

[10] A. Schwing, S. Fidler, M. Pollefeys, and R. Urtasun. Box in the box: Joint 3d layout and object reasoning from single images. In *ICCV*, 2013.

[11] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. In *CVPR*, 2017.

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016.

[13] X. Chen and A. Gupta. Spatial memory for context reasoning in object detection. In *ICCV*, 2017.

[14] X. Li, Z. Jie, W. Wang, C. Liu, J. Yang, X. Shen, Z. Lin, Q. Chen, S. Yan, and J. Feng. Foveanet: Perspective-aware urban scene parsing. In *ICCV*, pages 784–792, 2017.

[15] R. Garg, G. B. Kumar, Carneiro, and I. Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *ECCV*, 2016.

[16] T. Zhou, M. Brown, N. Snavely, and D. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017.

[17] R. Mahjourian, M. Wicke, and A. Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. *ArXiv e-prints*, February 2018.

[18] J. Yu, A. Harley, and K. Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In *ECCV Workshops*, 2016.

[19] A. Kendall and R. Cipolla. Geometric loss functions for camera pose regression with deep learning. In *CVPR*, 2017.

[20] H. Rhodin, J. Spörri, I. Katircioglu, V. Constantin, F. Meyer, E. Müller, M. Salzmann, and P. Fua. Learning monocular 3d human pose estimation from multi-view images. In *CVPR*, 2018.

[21] G. Kanizsa. Subjective contours. In *Scientific American*, 1976.

[22] A. Kar, S. Tulsiani, J. Carreira, and J. Malik. Amodal completion and size constancy in natural scenes. In *ICCV*, 2015.

[23] E. Hsiao and M. Hebert. Occlusion reasoning for object detection under arbitrary viewpoint. *CVPR*, 2012.

[24] S. Hochstein and M. Ahissar. View from the top: hierarchies and reverse hierarchies in the visual system. In *Neuron*, 2002.

[25] J. Huang and K. Murphy. Efficient inference in occlusion-aware generative models of images. In *ICLR Workshop*, 2016.

[26] L.L. Zhang, L. Lin, X.D. Liang, and K.M. He. Is faster r-cnn doing well for pedestrian detection? In *ECCV*, pages II: 443–457, 2016.

[27] R. Stewart, M. Andriluka, and A. Ng. End-to-end people detection in crowded scenes. In *CVPR*, 2016.

[28] H. Hattori, N. Lee, V. N. Boddeti, F. Beainy, K. M. Kitani, and T. Kanade. Synthesizing a scene-specific pedestrian detector and pose estimator for static video surveillance. 100(1), October 2018.

[29] P. Baque, F. Fleuret, and P. Fua. Deep occlusion reasoning for multi-camera multi-target detection. pages 271–279, 2017.

[30] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[31] A. Gupta, A. Mittal, and L. Davis. Constraint integration for multiview pose estimation of humans with self-occlusions. In *IEEE PAMI*, 2008.

[32] T. Chavdarova, P. Baqué, S. Bouquet, A. Maksai, C. Jose, T. Bagautdinov, L. Lettry, P. Fua, L. Gool, and F. Fleuret. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *CVPR*, 2018.

[33] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. Mot16: A benchmark for multi-object tracking. In *arXiv:1603.00831*, 2016.

[34] A. Shrivastava, A. Gupta, and R. Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, 2016.

Preprint submitted to 2019 International Conference on Robotics and Automation (ICRA). Received September 15, 2018.