Tracking Pedestrian Heads in Dense Crowd

Ramana Sundararaman   Cédric De Almeida Braga   Eric Marchand   Julien Pettré

Univ Rennes, Inria, CNRS, Irisa, Rennes, France

ramanasubramanyam.sundararaman@polytechnique.edu,

{cedric.de-almeida-braga,julien.pettre}@inria.fr,

eric.marchand@irisa.fr

Abstract

Tracking humans in crowded video sequences is an important constituent of visual scene understanding. Increasing crowd density challenges visibility of humans, limiting the scalability of existing pedestrian trackers to higher crowd densities. For that reason, we propose to revitalize head tracking with Crowd of Heads Dataset (CroHD), consisting of 9 sequences of 11,463 frames with over 2,276,838 heads and 5,230 tracks annotated in diverse scenes. For evaluation, we proposed a new metric, IDEucl, to measure an algorithm's efficacy in preserving a unique identity for the longest stretch in image coordinate space, thus building a correspondence between pedestrian crowd motion and the performance of a tracking algorithm. Moreover, we also propose a new head detector, HeadHunter, which is designed for small head detection in crowded scenes. We extend HeadHunter with a Particle Filter and a color histogram based re-identification module for head tracking. To establish this as a strong baseline, we compare our tracker with existing state-of-the-art pedestrian trackers on CroHD and demonstrate superiority, especially in identity preserving tracking metrics. With a light-weight head detector and a tracker which is efficient at identity preservation, we believe our contributions will serve useful in advancement of pedestrian tracking in dense crowds. We make our dataset, code and models publicly available at https://project.inria.fr/crowdscience/ project/dense-crowd-head-tracking/.

1. Introduction

Tracking multiple objects, especially humans, is a central problem in visual scene understanding. The intricacy of this task grows with increasing targets to be tracked and remains an open area of research. Alike other subfields in Computer Vision, with the advent of Deep Learning, the task of Multiple Object Tracking (MOT) has remarkably advanced its benchmarks [12, 24, 25, 37, 42, 61] since its
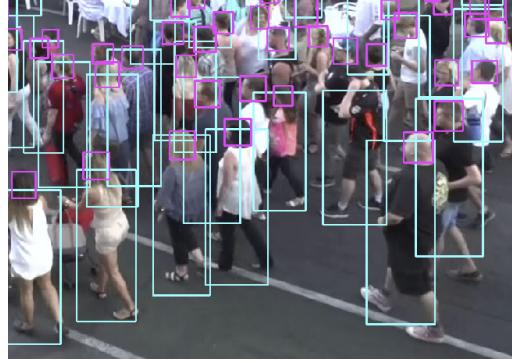

Figure 1. Comparison between head detection and full body detection in a crowded scene from CroHD. HeadHunter detects 36 heads whereas Faster-RCNN [48] can detect only 23 pedestrians out of 37 present in this scene.

inception [21]. In the recent past, the focus of MOTChallenge benchmark [13] has shifted towards tracking pedestrians in crowds of higher density. This has several applications in fields such as activity recognition, anomaly detection, robot navigation, visual surveillance, safety planning etc. Yet, the performances of trackers on these benchmark suggests a trend of saturation[1]. Majority of online tracking algorithms today follow the tracking-by-detection paradigm and several research works have well-established object detector's performance to be crucial in tracker's performance [3, 5, 11]. As the pedestrian density in a scene increases, pedestrian visibility reduces with increasing mutual occlusions, leading to reduced pedestrian detection as visualized in Figure 1. To tackle these challenges yet track humans efficiently in densely crowded environments, we rekindle the task of MOT with tracking humans by their distinctly visible part - heads. To that end, we propose a new dataset, *CroHD, Crowd of Heads Dataset*, comprising 9 sequences of 11,463 frames with head bounding boxes annotated for tracking. We hope that this new dataset opens

---

[1]https://motchallenge.net/results/MOT20/

up opportunities for promising future research to better understand global pedestrian motion in dense crowds.

Supplementing this, we develop two new baseline methods on CroHD, a head detector, *HeadHunter* and a head tracker, *HeadHunter-T*. We design HeadHunter peculiar for head detection in crowded environments, distinct from standard pedestrian detectors and demonstrate state-of-the-art performance on an existing head detection dataset. HeadHunter-T extends HeadHunter with a Particle Filter framework and a light-weight re-identification module for head-tracking. To validate HeadHunter-T to be a strong baseline tracker, we compare it with three published top performing pedestrian trackers on the crowded MOTChallenge benchmark, evaluated on CroHD. We further perform comparisons between tracking by head detection and tracking by body detection to illustrate the value of our contribution.

To establish correspondence between a tracking algorithm and pedestrian motion, it is necessary to understand the adequacy of various trackers in successfully representing ground truth pedestrian trajectories. We thus propose a new metric, *IDEucl* to evaluate tracking algorithms based on their consistency in maintaining the same identity for the longest length of a ground truth trajectory in the image coordinate space. *IDEucl* is compatible with our dataset and can be extended to any tracking benchmark, recorded with a static camera.

In summary, this paper makes the following contributions **(i)** We present a new dataset, CroHD, with annotated pedestrian heads for tracking in dense crowd, **(ii)** We propose a baseline head detector for CroHD, HeadHunter, **(iii)** We develop HeadHunter-T, by extending HeadHunter as the baseline head tracker for CroHD, **(iv)** We propose a new metric, IDEucl, to evaluate the efficiency of trackers in representing a ground truth trajectory and finally, **(v)** We demonstrate HeadHunter-T to be a strong baseline by comparing with three existing state-of-the-art trackers on CroHD.

2. Related Work

Head Detection Benchmarks: The earliest benchmarks in head detection are [29, 46, 62, 64], which provide ground truth head annotations of subjects in Hollywood movies. In the recent past, SCUT-Head [47] and CrowdHuman dataset [52] provide head annotations of humans in crowded scenes. Head detection is also of significant interest in the crowd counting and analysis literature [32]. Rodriguez *et al*. [50] introduced the idea of tracking by head detection with their dataset consisting of roughly 2200 head annotations. In the recent years, there has been a surge in research works attempting to narrow the gap between detection and crowd counting [39, 51, 41, 69] which attempts to hallucinate pseudo head ground truth bounding boxes in crowded scenes.

Head Detection Methods: Fundamentally, the task of

head detection is a combination of multi-scale and contextual object detection problem. Objects at multiple scales are detected based on image pyramids [30, 47, 55, 56, 66] or feature pyramids [26, 38, 70]. The former is computationally intensive task requiring multiple forward passes of images while the latter generates multiple pyramids in a single forward pass. Contextual object detection has been widely addressed in the literature of face detection, such as [14, 43, 60] who show improved detection accuracy by using convolutional filters of larger receptive size to model context. Sun *et al*. [58] employ such a contextual and scale-invariant applied to head detection.

Tracking Benchmarks and Metrics: The task of Multiple Object Tracking (MOT) is to track an initially unknown number of targets in a video sequence. The first MOT dataset for tracking humans were the PETS dataset [21], soon followed by [1, 16, 24, 25]. Standardization of MOT benchmarks were later proposed in [37] and since then, it has been updated with yearly challenges with increasingly crowded scenes [13, 42]. Recently, the TAO dataset [12] was introduced for Multi-object tracking, which focuses on tracking 833 object categories across 2907 short sequences. Our dataset pushes the challenge of tracking in crowded environments with pedestrian density reaching 346 humans per frame. Other relevant pedestrian tracking dataset include [8, 9, 61].

To evaluate algorithms on MOTChallenge dataset, classical MOT metrics [63] and CLEAR MOT metrics [4] have been *de facto* established as standardised way of quantifying performances. The CLEAR Metric proposes two important scores MOTA and MOTP which concisely summarise the classical metrics based on cumulative per frame accuracy and precision of bounding boxes respectively. Recently, Ristani *et al*. [49] propose the ID metric, which rewards a tracker based on its efficiency in preserving an identity for the longest duration of the Ground Truth trajectory.

Tracking Algorithms: Online Multi-object tracking algorithms can be summarised into: (i) Detection, (ii) Motion Prediction, (iii) Affinity Computation and, (iv) Association steps. R-CNN based networks have been common choice for the detection stage due to the innate advantage of proposal based detectors over Single-Stage detection methods [31]. Amongst online Multiple Object Tracking algorithm, Chen *et al*. [10] use Particle Filter framework and weigh the importance of each particle by their appearance classification score, computed by a separate network, trained independently. Earlier works such as [7, 33] use Sequential Importance Sampling (SIS) with Constant Velocity Assumption for assigning importance weight to particles. Henschel *et al*. [28] demonstrated the the limitation of single object detector for tracking and used a head detector [57] in tandem with the pedestrian detector [48]. However, in the recent past, research works in MOT have attempted to bridge

| Scene 1 | Scene 2 | Scene 3 | Scene 4 | Scene 5 |

Figure 2. Depiction of a frame per each scene from our Crowd of Heads Dataset, CroHD. The top row shows frames from the training set while the bottom row illustrates frames from the test set. Scene 5 is kept exclusive for the test set.

the gap between tracking and detection through a unified framework [3, 18, 19, 35, 40, 61]. Most notable amongst them is Tracktor [3], who demonstrated that an object detector alone is sufficient to predict locations of targets in subsequent frames, benefiting from the high-frame rates in video.

3. CroHD Dataset

Description: The objective of CroHD is to provide tracking annotation of pedestrian heads in densely populated video sequences. To the best of our knowledge, no such benchmark exists in the community and hence we annotated 2,276,838 human heads in 11,463 frames across 9 sequences of Full-HD resolution. We built CroHD upon 5 sequences from the publicly available MOTChallenge CVPR19 benchmark [13] to enable performance comparison of trackers in the same scene between two paradigms - head tracking and pedestrian tracking. We maintain the training set and test set classification of the aforementioned sequences to be the same in CroHD as the MOTChallenge CVPR19 benchmark. We further annotated 4 new sequences of higher crowd densities in two new scenarios. The new scenario centers on the Shibuya Train station and Shibuya Crossing, one of the busiest pedestrian crossings in the world. All sequences in CroHD have a framerate of $25 fps$ and are captured from an elevated viewpoint. The sequences involve crowded indoor and outdoor scenes, recorded across different lighting and environmental conditions. This ensures sufficient diversity in the dataset in order to make it viable for training and evaluating the comprehensiveness of modern Deep Learning based techniques. The maximum pedestrian density reaches approximately 346 persons per frame while the average pedestrian density across the dataset is 178. A detailed sequence-wise summary of CroHD is given in Table 1. We split CroHD into 4 sequences of 5740 frames for training and 5 sequences of

5723 frames for testing. They share three scenes in common, while the fourth scene is disparate to ensure generalization of trackers on this dataset. A representative frame from each sequence of CroHD and their respective training, testing splits are depicted in Figure 2. We will make our sequences and training set annotations publicly available. To preserve the fairness of the MOTChallenge CVPR19 benchmark, we will not release the annotations corresponding to the test set.

Annotation: The annotation and data format of CroHD follows the standard guidelines outlined by MOTChallenge benchmark [13, 42]. We annotated all visible heads of humans in a scene with the visibility left to the best of discretion of annotators. Heads of all humans, whose shoulder is visible were annotated, including the heads occluded by head coverings such as hood, caps etc. For sequences inherited from MOTChallenge CVPR19 benchmark, the annotations were performed independent of pedestrian tracking ground truth in order to have no dependencies between the two modalities. Due to the high frame rate in our video sequences, we interpolate annotations in between keyframes and adjust a track only when necessary.

CroHD constitutes four classes - Pedestrian, Person on Vehicle, Static and Ignore. Heads of statues or human faces on clothing have been annotated with an ignore label. Heads of pedestrians on vehicles, wheelchairs or baby transport have been annotated as Person on Vehicle. Pedestrians who do not move throughout the sequence are classified as static persons. Unlike the case of standard MOTChallenge benchmarks, we observe that overlap between bounding boxes are minimal since head bounding boxes from an elevated viewpoint are almost distinct. Hence, we limit our visibility flag to be binary - either visible (1.0) or occluded (0.0). We consider a proposal to be a match if the Intersection Over Union (IoU) with the ground truth is larger than 0.4.

Name	Frames	Scenario	Tracks	Boxes	Density
CroHD-01	429	Indoor	85	21,456	50.0
CroHD-02	3,315	Outdoor, night	1,276	733,622	222.0
CroHD-03	1,000	Outdoor, day	811	258,012	258.0
CroHD-04	997	Indoor	580	175,703	176.2
CroHD-11	584	Indoor	133	38,492	65.8
CroHD-12	2,080	Outdoor, night	737	383,677	185.0
CroHD-13	1,000	Outdoor, day	725	257,828	258.0
CroHD-14	1,050	Outdoor, day	562	258,227	246.0
CroHD-15	1,008	Outdoor, day	321	149,821	149.0
Total	11,463		5,230	2,276,838	178

Table 1. Sequence-wise statistics CroHD. Sequences are named CroHD-XY, with X being either 0 or 1 depending on training set or testing set respectively. Y denotes the serial number of videos.
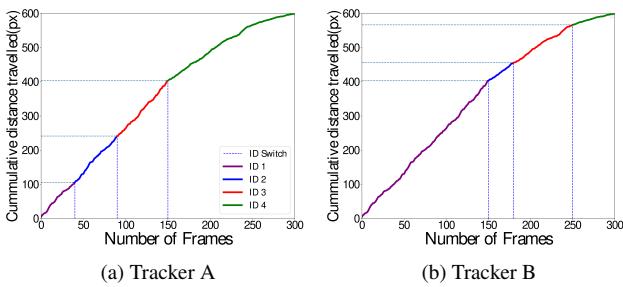

(a) Tracker A   (b) Tracker B

Figure 3. Identity prediction of two trackers - Tracker A (3a) and Tracker B (3b) for the same ground truth. A change of color implies an identity switch with both trackers registering 3 switches.

4. Evaluation Metrics

For evaluation of head detection on CroHD, we follow the standard Multiple Object detection metrics - mean Average Precision (mAP), Multiple Object Detection Accuracy (MODA), Multiple Object Detection Precision (MODP) [23] and mAP_COCO respectively. mAP_COCO refers to a stricter metric which computes the mean of AP across IoU thresholds of $\{50\%, 55\%, 60\%, \ldots, 95\%\}$. For evaluation of trackers, we adapt the well established Multiple Object Tracking metrics [4, 49], and extend with the proposed "IDEucl" metric.

IDEucl: While the event based metrics [4] and identity based metric (IDF1) [49] are persuasive performance indicators of a tracking algorithm from a local and global perspective, they do not quantify the proportion of the ground truth trajectory a tracker in capable of covering. Specifically, existing metrics do not measure the proportion of ground truth trajectory in the image coordinate space a tracker is able to preserve an identity. It is important to quantitatively distinguish between trackers which are more effective in tracking a larger portion of ground truth pedestrian trajectories. This is particularly useful in dense crowds, for better understanding of global crowd motion

pattern [15]. To that end, we propose a new evaluation metric, "IDEucl", which gauges a tracker based on its efficiency in maintaining consistent identity over the length of ground truth trajectory in image coordinate space. Albeit, IDEucl might seem related to the existing IDF1 metric which measures the fraction of frames of a ground truth trajectory in which consistent ID is maintained. In contrast, IDEucl measures the fraction of the distance travelled for which the correct ID is assigned.

To elucidate this difference, consider the example shown in Figure 3. Two trackers A and B compute four different identities for a ground truth trajectory G. Tracker A commits three identity switches in the first 150 frames while maintaining consistent identity for the remaining 150 frames. Tracker B, on the other hand, maintains consistent identity for the first 150 frames but commits 3 identity switches in the latter 150 frames. Our metric reports a score of 0.3 for Tracker A (Figure 3a) and a score of 0.67 for Tracker B (Figure 3b). Meanwhile, IDF1 and the classical metric reports a score of "0.5" and "3 identity switches" respectively for both the trackers. Following existing metrics, Tracker A and Tracker B are considered equally efficient. They neither highlight the ineffectiveness of Tracker A nor the ability of Tracker B in covering an adequate portion of ground truth trajectory with consistent identity. Therefore, IDEucl is more appropriate for judging the quality of the estimated pedestrian motion.

Thus, to formulate this metric, we perform a global hypothesis to ground truth matching by constructing a Bipartite Graph $\mathcal{G} = (\mathcal{U}, \mathcal{V}, \mathcal{E})$, similar to [49]. Two "regular" nodes are connected by an edge e if they overlap in time, with the overlap defined by Δ,

$$\Delta_{t,t-1} = \begin{cases} 1, & \text{if } \delta > 0.5 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Considering τ_t, h_t to be an arbitrary ground truth and hypothesis track at time t, δ is defined as,

$$\delta = \text{IoU}(\tau_t, h_t) \quad (2)$$

The cost on each edge $\mathcal{E} \in \mathbb{R}^N$ of this graph, $\mathcal{M} \in \mathbb{R}^{N-1}$ is represented as the distance in image space between two successive temporal associations of "regular" node. In particular, cost of an edge is defined as,

$$\mathcal{M} = \sum_{t=1}^{N} m_t = \begin{cases} \text{d}(\tau_t, \tau_{t-1}), & \text{if } \Delta_{t,t-1} = 1. \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where d denotes the Euclidean distance in image coordinate space. A ground truth trajectory is assigned a unique hypothesis which maintains a consistent identity for the predominant distance of ground truth in image coordinate space. We employ the Hungarian algorithm to solve

this maximum weight matching problem to obtain the best (longest) hypothesis. Once we obtain an optimal hypothesis, we formulate the metric $\mathcal{C}$ as the ratio of length of ground truth in image coordinates covered by the best hypothesis,

$$\mathcal{C} = \frac{\sum_{i=1}^{K} \mathcal{M}_i}{\sum_{i=1}^{K} \mathcal{T}_i} \qquad (4)$$

Note that this formulation of cost function naturally weighs the significance of each ground truth track based on its distance in image coordinate space.

5. Methods : Head Detection and Tracking

In this section, we elucidate the design and working of HeadHunter and HeadHunter-T.

5.1. HeadHunter

As detection is the pivotal step in object tracking, we designed HeadHunter differently from traditional object detectors [20, 48, 65] by taking into account the nature and size of objects we detect. HeadHunter is an end-to-end two stage detector, with three functional characteristics. First, it extracts feature at multiple scales using Feature Pyramid Network (FPN) [38] using a Resnet-50 [27] backbone. Images of heads are homogeneous in appearance and often, in crowded scenes, resemble extraneous objects (typically background). For that reason, inspired by the head detection literature, we augmented on top of each individual FPNs, a Context-sensitive Prediction Module (CPM) [60]. This contextual module consists of 4 Inception-ResNet-A blocks [59] with 128 and 256 filters for 3×3 convolution and 1024 filters for 1×1 convolution. As detecting pedestrian heads in crowded scenes is a problem of detecting many small-sized adjacently placed objects, we used Transpose Convolution on features across all pyramid levels to upscale the spatial resolution of each feature map. Finally, we used a Faster-RCNN head with Region Proposal Network (RPN) generating object proposals while the regression and classification head, each providing location offsets and confidence scores respectively. The architecture of our proposed network is summarised in Figure 4.

5.2. HeadHunter-T

We extended HeadHunter with two motion models and a color histogram based re-identification module for head-tracking. Our motion models consist of Particle Filter to predict motion of targets and Enhanced Correlation Coefficient Maximization [17] to compensate the Camera motion in the sequence. A Particle Filter is a Sequential Monte Carlo (SMC) process, which recursively estimates the state of dynamic systems. In our implementation, we represent

the posterior density function by a set of bounding box proposals for each target, referred to as particles. The use of Particle Filter enables us to simultaneously model nonlinearity in motion occurring due to rapid movements of heads and pedestrian displacement across frames.

Notation: Given a video sequence $\mathcal{I}$, we denote the ordered set of frames in it as $\{I_0, \cdots, I_{T-1}\}$, where T is the total number of frames in the sequence. Throughout the paper, we use subscript notation to represent time instance in a video sequence. In a frame I_t at time t, the active tracks are denoted by $\mathbf{T}_t = \{\mathbf{b}_t^1, \mathbf{b}_t^2, \ldots, \mathbf{b}_t^N\}$, where $\mathbf{b}_t^k$ refers to bounding box of the kth active track, denoted as $\mathbf{b}_t^k = (\mathbf{x_t^k}, \mathbf{y_t^k}, \mathbf{w_t^k}, \mathbf{h_t^k})$. At time t, the ith particle corresponding to k^{th} track is denoted by $\mathbf{p}_t^{k,i}$ and its respective importance weight by $\mathbf{w}_t^{k,i}$. $\mathbf{L}_t$ and $\mathbf{N}_t$ denote the set of inactive tracks and newly initialized tracks respectively.

Particle Initialization: New tracks are initialized at the start of the sequence, I_0 from the detection provided by HeadHunter and at frame I_t for detection(s) which cannot be associated with an existing track. A plausible association of new detection with existing track is resolved by Non-Maximal-Suppression (NMS). The importance weights of each particle are set to be equal at the time of initialisation. Each particles represent 4 dimensional state space, with the state of each targets modelled as $(\mathbf{x}_c, \mathbf{y}_c, \mathbf{w}, \mathbf{h}, \dot{\mathbf{x}}_c, \dot{\mathbf{y}}_c, \dot{\mathbf{w}}, \dot{\mathbf{h}})$, where, $(\mathbf{x}_c, \mathbf{y}_c, \mathbf{w}, \mathbf{h})$ denote the centroids, width and the height of bounding boxes.

Prediction and Update: At time $t > 0$, we perform RoI pooling on the current frame's feature map, $\mathbf{F}_t$, with the bounding box of particles corresponding to active tracks. Each particles' location in the current frame is then adjusted using the regression head of HeadHunter, given their location in the previous frame. The importance weights of each particle are set to their respective foreground classification score from the classification head of HeadHunter. Our prediction step is similar to the Tracktor [3], applied to particles instead of tracks. Given the new location and importance weight of each particle, estimated position of kth track is computed as weighted mean of the particles,

$$\mathbf{S}_t^k = \frac{1}{M} \sum_{i=1}^{M} \mathbf{w}_t^{k,i} \mathbf{p}_t^{k,i} \qquad (5)$$

Resampling: Particle Filtering frameworks are known to suffer from degeneracy problems [2] and as a result we resample to replace particles of low importance weight. M particles corresponding to k^{th} track are re-sampled when the number of particles which meaningfully contributes to probability distribution of location of each head, $\hat{\mathbf{N}}_{\text{eff}}^k$ exceeds a threshold, where,

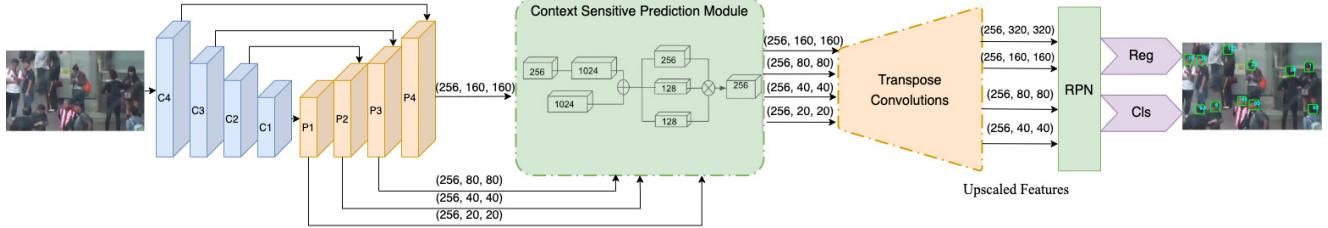$$\hat{\mathbf{N}}_{\text{eff}}^k = \frac{1}{\sum_{i=1}^{M} (\mathbf{w}^{k,i})^2} \qquad (6)$$

Figure 4. An overview of the architecture of our proposed head detector, Headhunter. We augment the features extracted using FPN (C4...P4) with Context Sensitive feature extractor followed by series of transpose convolutions to enhance spatial resolution of feature maps. Cls and Reg denote the Classification and Regression branches of Faster-RCNN [48] respectively.

Cost Matching: Tracks are set to inactive when scores of their estimated state $\mathbf{S}_t^a$ falls below a threshold, λ_{nms}^{reg}. Positions of such tracks are predicted following Constant Velocity Assumption (CVA) and their tracking is resumed if it has a convincing similarity with a newly detected track. The similarity, $\mathbf{C}$ is defined as

$$\mathbf{C} = \alpha \cdot IoU(\mathbf{L}_t^i, \mathbf{N}_t^j) + \beta \cdot d^1(\mathbf{L}_t^i, \mathbf{N}_t^j) \qquad (7)$$

where $\mathbf{L}_t^i$ and $\mathbf{N}_t^j$ are the ith lost track and j^{th} new track respectively. And, d^1 denotes the Bhattacharyya distance between the respective color histograms in the HSV space [45]. Once tracks are re-identified, we re-initialize particles around its new position.

6. Experiments

6.1. HeadHunter

We first detail the experimental setup and analyse the performance of HeadHunter on two datasets - SCUT-HEAD [47] and CroHD respectively. For the Faster-RCNN head of HeadHunter, we used 8 anchors, whose sizes were obtained by performing K-means over ground truth bounding boxes from the training set. To avoid overlapping anchors, they were split equally across the four pyramid levels, with the stride of anchors given by $\max(16, s/d)$ where s is square-root of the area of an anchor-box and d is the scaling factor [44]. For all experiments, we used Online Hard Example Mining [54] with 1000 proposals and a batch size of 512.

SCUT-Head is a large-scale head detection dataset consisting of 4405 images and 111,251 annotated heads split across Part A and Part B. We trained HeadHunter for 20 epochs with the input resolution to be the median image resolution of the training set (1000x600 pixels) and an initial learning rate of 0.01 halved at 5th, 10th and 15th epochs respectively. For a fair comparison, we trained HeadHunter only on the training set of this dataset and do not use any

pre-trained models. We summarize the quantitative comparisons with other head detectors on this dataset in Table 2. HeadHunter outperforms other state-of-the-art head detectors based on Precision, Recall and F1 scores.

Methods	Precision	Recall	F1
Faster-RCNN [48]	0.87	0.80	0.83
R-FCN+FRN [47]	0.91	0.84	0.87
SMD [58]	0.93	0.90	0.91
HSFA2Net [53]	0.94	0.92	0.93
HeadHunter (Ours)	**0.95**	**0.93**	**0.94**

Table 2. Comparison between HeadHunter's and other state-of-the-art head detectors on the SCUT-Head dataset.

CroHD: We first trained HeadHunter on the combination of training set images from SCUT-HEAD dataset and CrowdHuman dataset [52] for 20 epochs at a learning rate of 0.001. With variations well characterized, pre-training on large-scale image dataset improves the robustness of head detection. We then fine-tuned HeadHunter on the training set of CroHD, for a total of 25 epochs with an initial learning rate of 0.0001 using the ADAM optimizer [36]. The learning rate is then decreased by a factor of 0.1 at 10th and 20th epochs respectively.

Ablation: We examined our design choices for Head-Hunter, namely the use of context module and the anchor selection strategy by removing them. The head detection performance of HeadHunter and its variants on CroHD are summarised in Table 3. We threshold the minimum confidence of detection to 0.5 for evaluation. W/O Cont refers to the HeadHunter without Context Module. We further removed the median anchor sampling strategy and refer to as W/O Cont, mAn. We also provide baseline performance of Faster-RCNN with Resnet-50 backbone on CroHD, the object detector upon which we built HeadHunter. We followed the same training strategy for Faster-RCNN as Head-Hunter. All variants of HeadHunter significantly outperformed Faster-RCNN. Inclusion of the context module and

the anchor initialisation strategy also has a noteworthy impact on head detection.

Method	Precision	Recall	F1	MODA	MODP	mAP_COCO
Faster-RCNN [48]	34.4	42.2	50.1	40.3	30.8	11.2
W/O Cont, mAn	40.9	50.8	57.8	38.1	37.8	14.4
W/O Cont	44.3	57.8	64.5	40.0	42.7	15.0
HeadHunter	**52.8**	**63.4**	**68.3**	**50.0**	**47.0**	**19.7**

Table 3. Summary of various head detector's performances on the test set of CroHD.

6.2. HeadHunter-T

For the Particle Filtering framework, we used a maximum of N=100 particles for each object. The N particles were uniformly placed around the initial bounding box. To ensure that particles were not spread immoderately and were distinct enough, we sampled particles from a Uniform distribution whose lower and upper limit were $((x - 1.5w, y - 1.5h), (x + 1.5w, y + -1.5h))$ respectively. Where, x, y, w, h denote the centroid, width and height of the initial bounding box. For the color based re-identification, we used 16, 16 and 8 bins for the H, S and V channels respectively, where the brightness invariant Hue [22] was used instead of the standard Hue. α, β, which denotes the importance of IoU and color histogram matching, corresponding to Equation 7 were set to 0.8 and 0.2 respectively. We deactivated a track if it remained inactive for $\lambda^{age} = 25$ frames or if its motion prediction falls outside the image coordinates.

We evaluated three state-of-the-art trackers on CroHD, namely, SORT [5], V-IOU [6] and Tracktor [3] to compare with HeadHunter-T. We chose methods which do not require any tracking specific training, whose implementations have been made publicly available and are top-performing on the crowded MOTChallenge CVPR19 benchmark [13]. For a fair comparison, we performed all experiments with head detection provided by HeadHunter, thresholded to a minimum confidence of 0.6. SORT is an online tracker, which uses a Kalman Filter motion model and temporally associates detection based on IoU matching and Hungarian Algorithm. V_IOU associates detection based on IoU matching and employs visual information to reduce tracking inconsistencies due to missing detection. Parameters for V_IOU and SORT were set based on fine-tuning on the training set of CroHD, as discussed in the supplementary material. We evaluated two variants of Tracktor, with and without motion model. Tracktor+MM denotes the Tracktor extended with Camera Motion Compensation [17] and CVA for inactive tracks. For the two versions of Tracktor, we set tracking parameters similar to HeadHunter. Table 6.2 summarises the performance of aforementioned methods on the test set of CroHD. HeadHunter-T outperforms all the other methods, and furthermore demonstrates superiority in iden-

Method	MOTA ↑	IDEucl ↑	IDF1 ↑	MT ↑	ML ↓	ID Sw. ↓
SORT [5]	46.4	58.0	48.4	49	216	**649**
V_IOU [6]	53.4	34.3	35.4	80	182	1890
Tracktor [3]	58.9	31.8	38.5	125	117	3474
Tracktor+MM [3]	61.7	44.2	45.0	141	104	2186
HeadHunter-T	**63.6**	**60.3**	**57.1**	**146**	**93**	892

Table 4. **Main tracking result** comparing the performances of various state-of-the-art trackers and HeadHunter-T on the test set of CroHD. The direction of arrows indicate smaller or larger desired value for the metric.

tity preserved tracking. Although Tracktor [3] is similar to HeadHunter-T, there is a noticeable difference in its head tracking performance. We hypothesize the use of Particle Filter framework, which can handle arbitrary posteriors, as the reason for improvement. This claim is justified in the forthcoming section.

6.3. Ablation Experiments

HeadHunter-T: In this section, we analyse the design choices, in particular, the utility of re-identification module and Particle Filter of HeadHunter-T on the training set of CroHD. The results are summarised in Table 5. For variations in motion model, we removed the Particle Filter and used simple Camera Motion Compensation, denoted as HT w/o PF. We also experimented with a reduced number of particles initialized around the head, with n=10, denoted as HT + 10F. Introducing Particle Filter noticeably improved identity preserving scores (IDF1 and IDEucl) for HT + 10F. Further increasing the number of filters to 100 demonstrated the best performance. However, using more than 100 filters resulted in either duplicates or immoderate spreading, which are undesirable. We removed the re-identification module, to understand its influence, denoted as w/o ReID. Although color histogram is a modest image descriptor, yet it drastically reduced the number of identity switches and showed superior performance in identity preserving metrics - IDEucl, IDF1. We also experimented with α and β values corresponding to the importance of IoU and histogram matching (Equation 7). We set β to 0.8 and α to 0.2 and this configuration is denoted as HT + sReID. Surprisingly, we observed more identity switches and a slight decrease in performance across other tracking metrics. HeadHunter-T, our final model, outperformed all the other variants.

Choice of Filter: To further substantiate our choice of a multi-modal filter, we replaced the Particle Filter of HeadHunter-T with a Kalman Filter motion model [34]. While both Kalman Filter and Particle Filter are recursive state estimation algorithms, Kalman Filter assumes the system to be linear with Gaussian noise [2] while Particle Filter's multimodal posterior distribution enables it to model states of nonlinear systems. We replaced the Particle Filter with a four state Kalman Filter to model the

Method	MOTA ↑	IDEucl ↑	IDF1 ↑	MT ↑	ML ↓	ID Sw. ↓
HT w/o PF	60.6	40.1	43.9	200	102	3652
HT + 10F	63.3	58.2	56.3	214	98	1534
HT w/o ReID	59.5	57.7	57.5	**225**	**91**	1411
HT + sReID	59.1	57.8	58.3	**225**	**91**	1280
HT + KF	63.4	53.8	55.9	214	93	2451
HeadHunter-T	**64.0**	**61.5**	**58.5**	**225**	**91**	**1247**

Table 5. Illustration of ablation studies of HeadHunter-T (denoted as HT) on the training set of CroHD. The direction of arrows indicate small or large desired metric values.

inter-frame displacement of bounding boxes with CVA. The four states are x, y centroid coordinates, the height and aspect ratio of bounding boxes respectively, similar to the SORT [5]. The performance of this tracker, denoted as HT + KF is summarised in Table 5. HeadHunter-T with Particle Filter demonstrates superior performance than its Kalman Filter variant with respect to all the tracking metrics reported and in particular, we observe major improvement in-terms of IDEucl metric. Motion of heads along with the pedestrian displacement induces non-linearities in the position of bounding boxes. Although pedestrian motion in general is non-linear, this issue is exacerbated with the small size of head bounding boxes. Hence, using a multi-modal posterior state estimation is necessary to address the perceptible impact of non-linear motion. We remark this to be the reason behind improvement in performance while using a Particle Filter in comparison to the Kalman Filter.
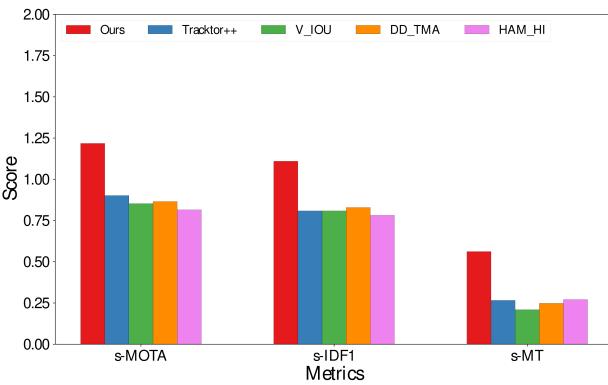

Figure 5. Comparison between HeadHunter-T and state-of-the-art trackers on common sequences of CroHD and MOTChallenge benchmark [13]. s-MOTA, s-IDF1, s-MT are scaled version of MOTA, IDF1 and Most Tracked (MT) metrics respectively.

Comparison across paradigm : We compare pedestrian and head tracking performances on the common sequences between CroHD and MOTChallenge CVPR19 dataset. The sequence being the same ensures that trackers are evaluated on full body and head bounding boxes

of the same pedestrians in the scene. For this comparison, we chose published state-of-the-art methods on the aforementioned dataset, namely, Tracktor++ [3], V_IOU, DD_TMA [68] and HAM_HI [67]. We performed comparison in-terms of MOTA, IDF1, MT (Mostly Tracked in percentage) metrics. Since we used a different object detector than the rest, a straightforward comparison between performance metrics would not be fair. Hence, for each sequence, we measure the ratio of aforementioned performance metrics with their object detector's MODA score to obtain the scaled scores - s-MOTA, s-IDF1 and s-MT. The scaled scores, averaged across five common sequences are illustrated in Figure 5. Our approach substantially outperforms other methods indicating that tracking by head detection is more suited for tracking in environments involving high pedestrian density where preserving identity is important. It is also worthy to note that HeadHunter uses a ResNet-50 backbone in contrast to a Resnet-101 backbone used by other methods. Furthermore, Tracktor++, HAM_HI and DD_TMA all use Deep Networks for extracting appearance features, while HeadHunter-T uses a color histogram based appearance feature. By compromising our tracking space (size of bounding box) to avoid mutual occlusion, we observe notable performance gain and significantly reduce the computation cost. This suggests that tracking by head detection paradigm is more desirable for real-time tracking applications focused on identity preservation.

7. Conclusion

To advance algorithms to track pedestrians in dense crowds, we introduced a new dataset, CroHD, for tracking by head detection. To further quantify the efficacy of a tracker in describing pedestrian motion, we introduced a new metric, IDEucl. We developed two new baseline methods, HeadHunter, HeadHunter-T for head detection and head tracking on CroHD respectively. We demonstrated HeadHunter-T to be consistently more reliable for identity preserving tracking applications than existing state-of-the-art trackers adapted for head tracking. Additionally, the adequacy of HeadHunter-T's performance with a modest computational complexity, opens up opportunities for future research focused on tracking methods adapted for low computational complexity and real-time applications. We also hope that CroHD will serve useful in contiguous fields, such as Crowd Counting and Crowd Motion Analysis.

References

[1] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[2] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, 2002.

[3] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixé. Tracking without bells and whistles. *CoRR*, abs/1903.05625, 2019.

[4] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008.

[5] Alex Bewley, ZongYuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. *CoRR*, abs/1602.00763, 2016.

[6] E. Bochinski, T. Senst, and T. Sikora. Extending iou based multi-object tracking by visual information. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2018.

[7] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool. Robust tracking-by-detection using a detector confidence particle filter. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1515–1522, 2009.

[8] Y. Cai and G. Medioni. Exploring context information for inter-camera multiple target tracking. In *IEEE Winter Conference on Applications of Computer Vision*, pages 761–768, 2014.

[9] Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. The WILDTRACK multi-camera person dataset. *CoRR*, abs/1707.09299, 2017.

[10] L. Chen, H. Ai, C. Shang, Z. Zhuang, and B. Bai. Online multi-object tracking with convolutional neural networks. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 645–649, 2017.

[11] Gioele Ciaparrone, Francisco Luque Sánchez, Siham Tabik, Luigi Troiano, Roberto Tagliaferri, and Francisco Herrera. Deep learning in video multi-object tracking: A survey. *Neurocomputing*, 381:61–88, Mar. 2020.

[12] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. Tao: A large-scale benchmark for tracking any object, 2020.

[13] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, and L. Leal-Taixé. CVPR19 tracking and detection challenge: How crowded can it get? *arXiv:1906.04567 [cs]*, June 2019. arXiv: 1906.04567.

[14] J. Deng, J. Guo, and S. Zafeiriou. Single-stage joint face detection and alignment. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 1836–1839, 2019.

[15] C. Dupont, L. Tobías, and B. Luvison. Crowd-11: A dataset for fine grained crowd behaviour analysis. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2184–2191, 2017.

[16] A. Ess, B. Leibe, K. Schindler, and L. van Gool. Moving obstacle detection in highly dynamic scenes. In *2009 IEEE International Conference on Robotics and Automation*, pages 56–63, 2009.

[17] G. D. Evangelidis and E. Z. Psarakis. Parametric image alignment using enhanced correlation coefficient maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1858–1865, 2008.

[18] Kuan Fang, Yu Xiang, and Silvio Savarese. Recurrent autoregressive networks for online multi-object tracking. *CoRR*, abs/1711.02741, 2017.

[19] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. *CoRR*, abs/1710.03958, 2017.

[20] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[21] J. Ferryman and A. Shahrokni. Pets2009: Dataset and challenge. In *2009 Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-Winter 2009)*, Los Alamitos, CA, USA, dec 2009. IEEE Computer Society.

[22] Graham Finlayson and Gerald Schaefer. Hue that is invariant to brightness and gamma. In *Proc. British Machine Vision Conference*, pages 303–312, 2001.

[23] John S. Garofolo, Rachel Bowers, Dennis E. Moellman, Rangachar Kasturi, Dmitry B. Goldgof, and Padmanabhan Soundararajan. Performance evaluation protocol for face, person and vehicle detection & tracking in video analysis and content extraction (vace-ii) clear - classification of events, activities and relationships. 2006.

[24] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32:1231 – 1237, 2013.

[25] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. pages 3354–3361, 05 2012.

[26] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[28] Roberto Henschel, Laura Leal-Taixé, Daniel Cremers, and Bodo Rosenhahn. Improvements to frank-wolfe optimization for multi-detector multi-object tracking. *CoRR*, abs/1705.08314, 2017.

[29] M. Hoai and A. Zisserman. Talking heads: Detecting humans and recognizing their interactions. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 875–882, 2014.

[30] Peiyun Hu and Deva Ramanan. Finding tiny faces. *CoRR*, abs/1612.04402, 2016.

[31] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, and Kevin Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. *CoRR*, abs/1611.10012, 2016.

[32] H. Idrees, K. Soomro, and M. Shah. Detecting humans in dense crowds using locally-consistent scale prior and global occlusion reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(10):1986–1998, 2015.

[33] Junliang Xing, H. Ai, and S. Lao. Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1200–1207, 2009.

[34] R. Kalman. A new approach to linear filtering and prediction problems" transaction of the asme journal of basic. 1960.

[35] H. Kieritz, W. Hübner, and M. Arens. Joint detection and online multi-object tracking. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1540–15408, 2018.

[36] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

[37] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler. MOTChallenge 2015: Towards a benchmark for multi-target tracking. *arXiv:1504.01942 [cs]*, Apr. 2015. arXiv: 1504.01942.

[38] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. *CoRR*, abs/1612.03144, 2016.

[39] Jiang Liu, Chenqiang Gao, Deyu Meng, and Alexander G. Hauptmann. Decidenet: Counting varying density crowds through attention guided detection and density estimation. *CoRR*, abs/1712.06679, 2017.

[40] Mason Liu, Menglong Zhu, Marie White, Yinxiao Li, and Dmitry Kalenichenko. Looking fast and slow: Memory-guided mobile video object detection. *CoRR*, abs/1903.10172, 2019.

[41] Yuting Liu, Miaojing Shi, Qijun Zhao, and Xiaofang Wang. Point in, box out: Beyond counting persons in crowds. *CoRR*, abs/1904.01333, 2019.

[42] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. MOT16: A benchmark for multi-object tracking. *arXiv:1603.00831 [cs]*, Mar. 2016. arXiv: 1603.00831.

[43] Mahyar Najibi, Pouya Samangouei, Rama Chellappa, and Larry S. Davis. SSH: single stage headless face detector. *CoRR*, abs/1708.03979, 2017.

[44] Mahyar Najibi, Bharat Singh, and Larry Davis. Fa-rpn: Floating region proposals for face detection. pages 7715–7724, 06 2019.

[45] Katja Nummiaro, Esther Koller-Meier, and Luc J. Van Gool. Object tracking with an adaptive color-based particle filter. In *Proceedings of the 24th DAGM Symposium on Pattern Recognition*, page 353–360, Berlin, Heidelberg, 2002. Springer-Verlag.

[46] A. Patron-Perez, M. Marszalek, I. Reid, and A. Zisserman. Structured learning of human interactions in tv shows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(12):2441–2453, 2012.

[47] Dezhi Peng, Zikai Sun, Zirong Chen, Zirui Cai, Lele Xie, and Lianwen Jin. Detecting heads using feature refine net and cascaded multi-scale architecture. *CoRR*, abs/1803.09256, 2018.

[48] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.

[49] Ergys Ristani, Francesco Solera, Roger S. Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. *CoRR*, abs/1609.01775, 2016.

[50] M. Rodriguez, J. Sivic, I. Laptev, and J. Audibert. Data-driven crowd analysis in videos. In *2011 International Conference on Computer Vision*, pages 1235–1242, 2011.

[51] Deepak Babu Sam, Skand Vishwanath Peri, Mukuntha Narayanan Sundararaman, Amogh Kamath, and Venkatesh Babu Radhakrishnan. Locate, size and count: Accurately resolving people in dense crowds via detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020.

[52] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018.

[53] Wenxiang Shen, Pinle Qin, and Jianchao Zeng. An indoor crowd detection network framework based on feature aggregation module and hybrid attention selection module. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 82–90. IEEE, 2019.

[54] Abhinav Shrivastava, Abhinav Gupta, and Ross B. Girshick. Training region-based object detectors with online hard example mining. *CoRR*, abs/1604.03540, 2016.

[55] Bharat Singh and Larry S Davis. An analysis of scale invariance in object detection-snip. *CVPR*, 2018.

[56] Bharat Singh, Mahyar Najibi, and Larry S Davis. SNIPER: Efficient multi-scale training. *NeurIPS*, 2018.

[57] Russell Stewart and Mykhaylo Andriluka. End-to-end people detection in crowded scenes. *CoRR*, abs/1506.04878, 2015.

[58] Z. Sun, D. Peng, Z. Cai, Z. Chen, and L. Jin. Scale mapping and dynamic re-detecting in dense head detection. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1902–1906, 2018.

[59] Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, abs/1602.07261, 2016.

[60] Xu Tang, Daniel K. Du, Zeqiang He, and Jingtuo Liu. Pyramidbox: A context-assisted single shot face detector. *CoRR*, abs/1803.07737, 2018.

[61] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. MOTS: multi-object tracking and segmentation. *CoRR*, abs/1902.03604, 2019.

[62] Tuan-Hung Vu, Anton Osokin, and Ivan Laptev. Context-aware CNNs for person head detection. In *International Conference on Computer Vision (ICCV)*, 2015.

[63] Bo Wu and Ramakant Nevatia. Tracking of multiple, partially occluded humans based on static body part detection. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 1:951–958, 2006.

[64] Xiaofeng Ren. Finding people in archive films through tracking. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[65] F. Yang, W. Choi, and Y. Lin. Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2129–2137, 2016.

[66] Young Joon Yoo, Dongyoon Han, and Sangdoo Yun. EXTD: extremely tiny face detector via iterative filter reuse. *CoRR*, abs/1906.06579, 2019.

[67] Y. Yoon, A. Boragule, Y. Song, K. Yoon, and M. Jeon. Online multi-object tracking with historical appearance matching and scene adaptive detection filtering. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2018.

[68] Young-Chul Yoon, Du Yong Kim, Young-Min Song, Kwangjin Yoon, and Moongu Jeon. Online multiple pedestrians tracking using deep temporal appearance matching association. *Information Sciences*, Oct. 2020.

[69] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma. Single-image crowd counting via multi-column convolutional neural network. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 589–597, 2016.

[70] Chenchen Zhu, Ran Tao, Khoa Luu, and Marios Savvides. Seeing small faces from robust anchor's perspective. *CoRR*, abs/1802.09058, 2018.