Learning to Detect and Track Visible and Occluded Body Joints in a Virtual World

Matteo Fabbri Fabio Lanzi Simone Calderara Andrea Palazzi
Roberto Vezzani Rita Cucchiara

University of Modena and Reggio Emilia, Italy

Abstract. Multi-People Tracking in an open-world setting requires a special effort in precise detection. Moreover, temporal continuity in the detection phase gains more importance when scene cluttering introduces the challenging problems of occluded targets. For the purpose, we propose a deep network architecture that jointly extracts people body parts and associates them across short temporal spans. Our model explicitly deals with occluded body parts, by hallucinating plausible solutions of not visible joints. We propose a new end-to-end architecture composed by four branches (*visible heatmaps*, *occluded heatmaps*, *part affinity fields* and *temporal affinity fields*) fed by a *time linker* feature extractor. To overcome the lack of surveillance data with tracking, body part and occlusion annotations we created the vastest Computer Graphics dataset for people tracking in urban scenarios by exploiting a photorealistic videogame. It is up to now the vastest dataset (about 500.000 frames, more than 10 million body poses) of human body parts for people tracking in urban scenarios. Our architecture trained on virtual data exhibits good generalization capabilities also on public real tracking benchmarks, when image resolution and sharpness are high enough, producing reliable tracklets useful for further batch data association or re-id modules.

Keywords: pose estimation, tracking, surveillance, occlusions

1 Introduction

Multi-People Tracking (MPT) is one of the most established fields in computer vision. It has been recently fostered by the availability of comprehensive public benchmarks and data [1,2]. Often, MPT approaches have been casted in the *tracking by detection paradigm* where a pedestrian detector extracts candidate objects and a further association mechanism arranges them in a temporally consistent trajectory [3,4,5]. Nevertheless, in the last years several researchers [6,3] raised the question on whether these two phases would be disentangled or considered two sides of the same problem. The strong influence between detection accuracy and tracking performance [3] suggests considering detection and tracking as two parts of a unique problem that should be addressed end-to-end at least for short-term setups. In this work, we advocate for an integrated approach between detection and short-term tracking that can serve as a proxy for more

complex association method either belonging to the tracking or re-id family of techniques. To this aim, we propose:

- an end-to-end deep network, called *THOPA-net (Temporal Heatmaps and Occlusions based body Part Association)* that jointly locates people body parts and associates them across short temporal spans. This is achievable with modern deep learning architectures that exhibit terrific performance in body part location [7] but, mostly, neglect the temporal contribution. For the purpose, we propose a bottom-up human pose estimation network with a temporal coherency module that jointly enhances the detection accuracy and allows for short-term tracking;
- an explicit method for dealing with occluded body parts that exploits the capability of deep networks of hallucinating feasible solutions;
- a massive computer graphics dataset, namely JTA (Joints Tracking Annotated dataset), that simulates realistic people tracking scenarios in a virtual world, in accordance with recent literature that testifies the advantage of disposing of virtual world proxy for several deep learning problems [8,9]. Our dataset is the first of its kind for people surveillance in urban scenarios and comes with a rich automatic annotation on people body part locations and their per-frame tracking. The dataset is composed by about 500K frames and 128 different scenarios from both fixed and moving cameras and it covers the most frequent challenges of MPT in urban areas: almost 20K identities, crowded scenes with up to 60 people, 10 millions of body poses.

Results are very encouraging in their precision also in crowded scenes. Our experiments tell us that the problem is less dependent on the details or the realism of the shape than one could imagine; instead, it is more affected by the image quality and resolution that are extremely high in Computer Graphics (CG) generated datasets. Nevertheless, experiments on real MPT dataset [1] demonstrate that with a minimal amount of fine-tuning the model can transfer positively towards real scenarios.

2 Related Works

Human pose estimation in images has made important progress over the last few years [10,11,12,13,14]. However, those techniques assume only one person per image and are not suitable for videos of multiple people that occlude each other. The natural extension of single-person pose estimation, i.e, multi-person pose estimation, has therefore gained much importance recently being able of handling situations with a varying number of people [15,16,17,18,19,7,20]. Among them, [18] uses graph decomposition and node labeling with local search while [19] introduces associative embeddings to simultaneously generate and group body joints detections. An end-to-end architecture for jointly learning body parts and their association is proposed by [7] while [18], instead, exploits a two-stage approach, consisting of a person detection stage followed by a keypoint estimation

for each person. Moreover, [15,16,17] jointly estimate multiple poses in the image, while also handling truncations and occlusions. However, those methods still rely on a separate people detector and do not perform well in cluttered situations.

Single person pose estimation in videos has been addressed by several researchers, [21,22,23,24]. Nevertheless, all those methods improve the pose estimation accuracy by exploiting temporal smoothing constraints or optical flow data, but neglect the case of multiple overlapping people.

In recent years, online tracking has been successfully extended to scenarios with multiple targets [25,26,27,28,29,30]. In contrast to single target tracking approaches, which rely on sophisticated appearance models to track a single entity in subsequent frames, multiple target tracking does not rely solely on appearance models. [25] exploits a high-performance detector with a deep learning appearance feature while [27] presents an online method that encodes long-term temporal dependencies across multiple cues. [28], on the other hand, introduces spatial-temporal attention mechanism to handle the drift caused by occlusion and interaction among targets. [29] solves the online multi-object tracking problem by associating tracklets and detections in different ways according to their confidence values and [30] exploits both high and low confidence target detections in a probability hypothesis density particle filter framework.

In this work, we address the problem of multi-person pose estimation in videos jointly with the goal of multiple people tracking. Early works that approach the problem [31,32] do not tackle pose estimation and tracking simultaneously, but rather target on multi-person tracking alone. More recent methods [33,34], which rely on graph partitioning approaches closely related to [15,16,17], simultaneously estimate the pose of multiple people and track them over time but do not cope with urban scenarios that are dominated by targets occlusions, scene clutterness and scale variations. In contrast to [33,34] we do not tackle the problem as a graph partitioning approach. Instead, we aim at simplifying the tracking problem by providing accurate detections robust to occlusions by reasoning directly at video level.

The most widely used publicly available datasets for human pose estimation in videos are presented in Tab. 1. [35,36,37] provide annotations for the single-person subtask of person pose estimation. Only Posetrack [2] has a multi-person perspective with tracking annotations but not provide them in the surveillance context. The reference benchmark for evaluation of multi-person tracking is [38] which provides challenging sequences of crowded urban scenes with severe occlusions and scale variations. However, it pursuits no pose estimation task and only provides bounding boxes as annotations. Our virtual world dataset instead, aim at taking the best of both worlds by merging precise pose and tracking annotations in realistic urban scenarios. This is indeed feasible when the ground truth can be automatically computed exploiting highly photorealistic CG environments.

Table 1. Overview of the publicly available datasets for Pose Estimation and MPT in videos. For each dataset we reported the numbers of clips, annotated frames and people per frame, as well as the availability of 3D data, occlusion labels, tracking information, pose estimation annotations and data type

Dataset	#Clips	#Frames	#PpF	3D	Occl.	Tracking	Pose Est.	Type
Penn Action [35]	2,326	159,633	1				✓	sports
JHMDB [36]	5,100	31,838	1				✓	diverse
YouTube Pose [37]	50	5,000	1				✓	diverse
Video Pose 2.0 [39]	44	1,286	1				✓	diverse
Posetrack [2]	514	23,000	1-13			✓	✓	diverse
MOT-16 [38]	14	11,235	6-51		✓	✓		urban
JTA	512	460,800	0-60	✓	✓	✓	✓	urban

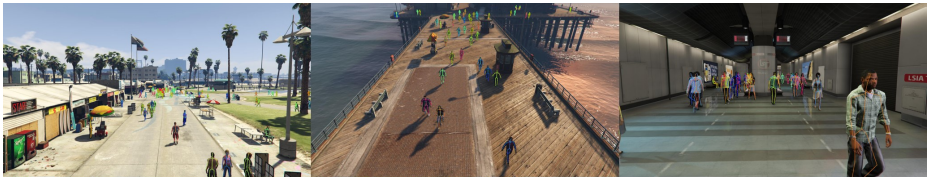

Fig. 1. Examples from the JTA dataset exhibiting its variety in viewpoints, number of people and scenarios. Ground truth joints are superimposed to the original images

3 JTA Dataset

We collected a massive dataset JTA for pedestrian pose estimation and tracking in urban scenarios by exploiting the highly photorealistic video game *Grand Theft Auto V* developed by *Rockstar North*. The collected videos feature a vast number of different body poses, in several urban scenarios at varying illumination conditions and viewpoints, Figure 1. Moreover, every clip comes with a precise annotation of visible and occluded body parts, people tracking with 2D and 3D coordinates in the game virtual world. In terms of completeness, our JTA dataset overcomes all the limitation of existing dataset in terms of number of entities and available annotations, Table 1. In order to virtually recreate real-world scenarios we manually directed the scenes by developing a game modification that interacts synchronously with the video game's engine[1]. The developed module allowed us to generate and record natural pedestrian flows recreating people behaviors specific to the most crowded areas. Moreover, exploiting the game's APIs, the software can handle people actions: in clips, people occasionally perform natural actions like sitting, running, chatting, talking on the phone, drinking or smoking. Each video contains a number of people ranging between 0 and 60 with an average of more than 21 people, totaling over 10M annotated body poses over 460,800 densely annotated frames. The distance from the camera ranges between 0.1 and 100 meters as depicted in Figure 2, resulting

[1] The mod will be publicly available upon publication.

in pedestrian heights between 20 and 1100 pixels. We collected a set of 512 Full

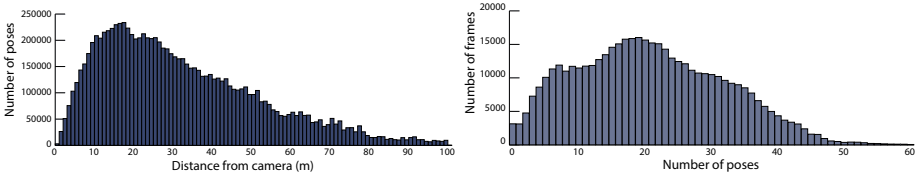

Fig. 2. Statistics about the JTA Dataset. (a) Number of annotated pose instances per camera distance (in meters). (b) Number of frames vs. the number of annotated poses per frame

HD videos, 30 seconds long, recorded at 30 fps. We halve the sequences into 256 videos for training and 256 for testing purposes. Through the game modification, we access the game renderer for automatically annotating the same 14 body parts in [40] and [2] in order to foster cross-dataset experiments. In each video, we assigned a unique identifier to every pedestrian that appears in the scene. The identifier remains the same throughout the entire video even if the pedestrian moves out the field-of-view. This feature could foster person re-identification research despite not being the target of this work. Our dataset also provides *occlusion* and *self-occlusion* flags. Each joint is marked as occluded if it is not directly visible from the camera point of view and it is occluded by objects or other pedestrians. Instead, a joint is marked as self-occluded if it is occluded by the same person to whom the joint belongs. As for joints annotation, occlusion annotation is captured by accessing the game renderer. JTA Dataset also provides accurate 3D information: for each annotated joint, as well as having the 2D coordinates of the location in the image, we also provide the 3D coordinates of the location in the simulator's space. Differently from Posetrack [2], which uses the annotated head bounding boxes as an estimation of the absolute scale of the person, we provide the precise scale of each pedestrian through the 3D annotation.

4 THOPA-net

Our approach exploits both intra-frame and inter-frame information in order to jointly solve the problem of multi-person pose estimation and tracking in videos. For individual frames, we extended the architecture in [7] by integrating a branch for handling occluded joints in the detection process. Subsequently, we propose a temporal linking network to integrate temporal consistency in the process and jointly achieve detection and short-term tracking. The Single Image model, Figure 3, takes an RGB frame of size $w \times h$ as input and produces, as output, the pose prediction for every person in the image. Conversely the

complete architecture, Figure 4, takes a clip of N frames as input and outputs the pose prediction for the last frame of the clip and the temporal links with previous frames.

4.1 Single Image Pose Prediction

Our single image model, Figure 3, consists of an initial feature extractor based on the first 10 layers of VGG-19 [41] pretrained on COCO 2016 keypoints dataset [42]. The computed feature maps are subsequently processed by a three-branch multi-stage CNN where each branch focuses on a different aspect of body pose estimation: the first branch predicts J heatmaps of the visible parts, the second branch predicts J heatmaps of the occluded parts and the third branch predicts C part affinity fields (PAFs), which are vector fields used to link parts together. Note that, oppositely to [7], we employed a different branch for the occlusion detection task. It is straightforward that visible and occluded body parts detection are two related but distinct tasks. The features used by the network in order to detect the location of a body part are different from those needed to estimate the location of an occluded one. Nevertheless, the two problems are entangled together since visible parts allow to estimate the missing ones. In fact, the network exploits contextual cues in order to perform the desired prediction, and the presence of a joint is indeed strongly influenced by the person's silhouette (e.g. a foot detection mechanism relies heavily on the presence of a leg, thus a visible foot detection may trigger even though the foot is not completely visible). Each branch is, in turn, an iterative predictor that refines the predictions at each subsequent stage applying intermediate supervision in order to address the vanishing gradient problem. Apart from the first stage, which takes as input only the features provided by VGG-19, the consecutive stages integrate the same features with the predictions from the branches at the previous stage. Consequently, information flow across the different branches and in particular both visible and occluded joints detection are entangled in the process.
We apply, for each branch, a different loss function at the end of each stage. The loss is a SSE loss between estimated predictions and ground truth, masked by a mask M in order to not penalize occluded joints in the visible branch. Specifically, for the generic output of each branch X^s of stage $s \in \{1, \dots, S\}$ and the ground truth X^* we have the loss function:

$$l_X^s = \sum_i \sum_{x=1}^{w'} \sum_{y=1}^{h'} M(x,y) \odot (X_i^s(x,y) - X_i^*(x,y))^2, \tag{1}$$

where X is in turn H for visible joints heatmaps, O for occluded ones and P for affinity fields; the outer summation spans the J number of joints for H and O and the C number of limbs for P. H^s, O^s and P^s sizes (w', h') are eight times smaller than the input due to VGG19 max pooling operations. Eventually, the overall objective becomes $L = \sum_{s=1}^{S} (l_H^s + l_O^s + l_P^s)$.

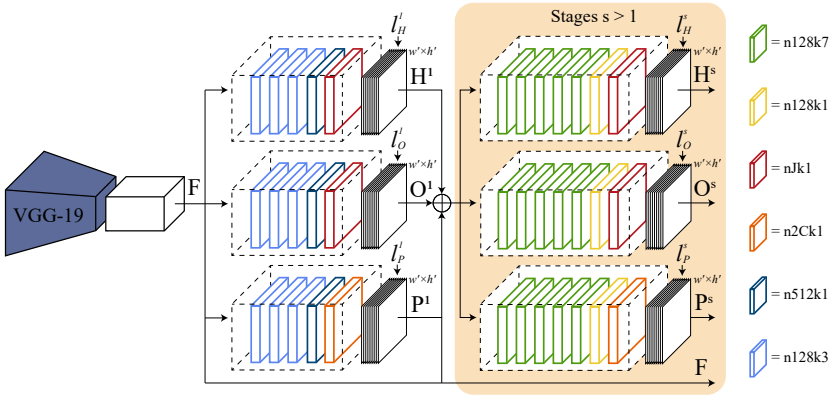

Fig. 3. Architecture of the three-branch multi-stage CNN with corresponding kernel size (k) and number of feature maps (n) indicated for each convolutional layer

4.2 Temporal Consistency Branch

In order to jointly solve the problem of multi-person pose estimation and tracking we enhance the Single Image model by adding our novel temporal network, Figure 4. The temporal model takes as input N RGB frames of size $w \times h$ and produces, as output, the temporal affinity fields (TAFs), as well as heatmaps and part affinity fields. TAFs, like PAFs, are vector fields that link body parts but oppositely to PAFs are focused on temporal links instead of spatial ones. In detail, PAFs connect different types of body parts intra-frame while TAFs, instead, connect the same types of body parts inter-frame, e.g, they connect heads belonging to the same person in two subsequent frames. The TAF field is, in fact, a proxy of the motion of the body parts and provide the expected location of the same body part in the previous frame and can be used both for boosting the body parts detection and for associating body parts detections in time. At a given time t_0, our architecture takes frames $I^t \in \mathbb{R}^{w \times h \times 3}$ with $t \in \{t_0, t_{-\tau}, t_{-2\tau}, \ldots, t_{-N\tau+1}\}$ and pushes them through the VGG19 feature extractor, described in Section 4.1, to obtain N feature tensors $f^t \in \mathbb{R}^{w' \times h' \times r}$ where r is the number of channels of the feature tensor. Those tensors are then concatenated over the temporal dimension obtaining $F \in \mathbb{R}^{w' \times h' \times r \times N}$. F is consecutively fed to a cascade of 3D convolution blocks that, in turn, capture the temporal patterns of the body part features and distill them by temporal max pooling until we achieve a feature tensor $F' \in \mathbb{R}^{w' \times h' \times r}$, Figure 4. As in Section 4.1, the feature maps are passed through a multi-branch multi-stage CNN. Moreover, we add to the Single Image architecture a fourth branch for handling the TAFs prediction. As a consequence, after the first stage, temporal information flow to all the branches of the network and acts as a prior for body part estimation (visible and occluded) and PAFs computation. The complete

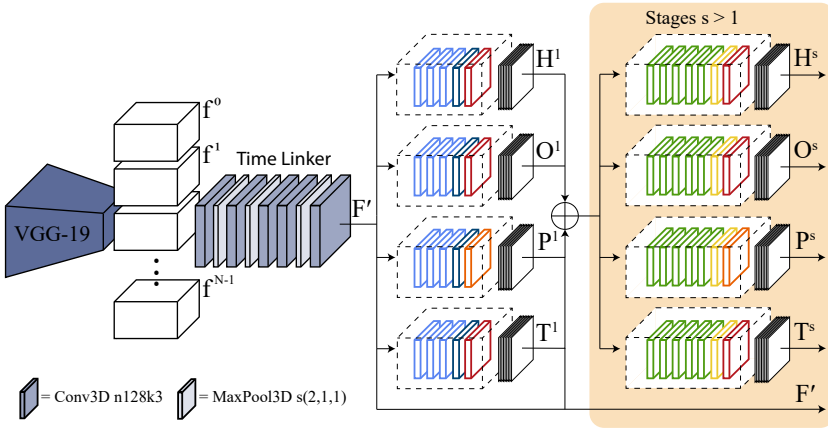

Fig. 4. Architecture of our method that encompass pose estimation and tracking in an end-to-end fashion. The MaxPool3D perform pooling operations only in the temporal dimension with stride s

network objective function then becomes $L = \sum_{s=1}^{S}(l_H^s + l_O^s + l_P^s + l_T^s)$ where

$$l_T^s = \sum_{j=1}^{J} \sum_{x=1}^{w'} \sum_{y=1}^{h'} M(x,y) \odot (T_j^s(x,y) - T_j^*(x,y))^2 \tag{2}$$

is the loss function computed between the ground truth T_j^* and the prediction T_j^s at each stage s. The set $T = (T_1, T_2, \ldots, T_j)$ has J vector fields, one for each part, with $T_j \in \mathbb{R}^{w \times h}, j \in \{1, \ldots, J\}$.

4.3 Training Procedure

During training, we generate both the ground truth heatmaps H^* and O^* from the annotated keypoint coordinates by placing at the keypoint location a 2D Gaussian with its variance conditioned by the true metric distance, d, of the keypoint from the camera. Oppositely to [7], by smoothing the Gaussian using distances, it is possible to achieve heatmaps of different sizes proportional to the scale of the person itself. This process is of particular importance to force scale awareness in the network and avoiding the need of multi scale branches. For example, given a visible heatmap H_j, let $q_{j,k} \in \mathbb{R}^2$ be the ground truth location of the body part j of the person k. For each body part j the ground truth H_j^* at location $p \in \mathbb{R}^2$ results:

$$H_j^*(p) = \max_k \exp\left(-\frac{\|p - q_{j,k}\|_2^2}{\sigma^2}\right), \qquad \sigma = \exp\left(1 - \frac{d}{\alpha}\right) \tag{3}$$

where σ regulates the spread of the peak in function of the distance d of each joint from the camera. In our experiments we choose α equals to 20.

Instead, each location p of ground truth part affinity fields $P^*_{c,k}$ is equal to the unit vector (with the same direction of the limb) if the point p belongs to the limb. The points belonging to the limb are those within a distance threshold of the line segment that connect the pair of body parts. For each frame, the ground truth part affinity fields are the two channels image containing the average of the PAFs of all people. As previously stated, by extending the concept of PAFs to the temporal dimension, we propose the novel TAFs representation which encodes short-term tubes of body parts across multiple frames (as shown in Figure 5.(b,d)). The temporal affinity field is a 2D vector field, for each body part, that points to the location of the same body part in the previous frame. Consider a body part j of a person k at frame t and let $q^{t-1}_{j,k}$ and $q^{t}_{j,k}$ be their ground truth positions at frame $t-1$ and t respectively. If a point p lies on the path crossed by the body part j between $t-1$ and t, the value at $T^*_{j,k}(p)$ is a unit vector pointing from j at time t to j at time $t-1$; for all other points the vector is zero. We computed ground truth TAFs using the same strategy exploited for PAFs.

(a) (b) (c) (d)

Fig. 5. (a) Annotations of visible (green), occluded (red) and self-occluded (orange) joints. (b) Ground truth examples of heatmaps, PAFs and TAFs. (c) Pose prediction performed on JTA dataset which distinguish between visible and occluded joints. (d) Head TAFs prediction on JTA dataset: the color encode the direction of the movement extrapolated from the previous frame

4.4 Spatio-Temporal Multi-Person Joints Association

In order to connect body parts into skeletons we take into account two different contributions both at frame level (PAF) and at temporal level (TAF). First, the joints heatmaps are non-maxima suppressed to obtain a set of discrete locations, D_j, for multiple people, where $D_j = \{d^m_j : \text{for } j \in \{1,\dots,J\}, m \in \{1,\dots,N_j\}\}$ and N_j is the number of candidates of part j, and J the number of joint types. We associate joints by defining a variable $z^{mn}_{j_1 j_2} \in \{0,1\}$ to indicate whether two joints candidates $d^m_{j_1}$ and $d^n_{j_2}$ are connected. Consequently, the objective is to find the optimal assignment for the set of possible connections, $Z = \{z^{mn}_{j_1 j_2} : \text{for } j_1, j_2 \in \{1,\dots,J\}, m \in \{1,\dots,N_{j_1}\}, n \in \{1,\dots,N_{j_2}\}\}$. To this aim we score

every candidate limb (i.e. a pair of joints) spatially and temporally by computing the line integral along PAFs, E and TAFs, G:

$$E(d_{j_1}, d_{j_2}) = \int_{u=0}^{u=1} PAF(p(u)) \cdot \frac{d_{j_2} - d_{j_1}}{\left\| d_{j_2} - d_{j_1} \right\|_2} du \tag{4}$$

$$G(d_j, \hat{d}_j) = \int_{u=0}^{u=1} TAF(t(u)) \cdot \frac{\hat{d}_j - d_j}{\left\| \hat{d}_j - d_j \right\|_2} du \tag{5}$$

where $p(u)$ linearly interpolates the locations along the line connecting two joints d_{j_2} and d_{j_1} and $t(u)$ acts analogously for two joints $\hat{d}_j$ at frame $t-1$ and d_j at frame t.
We then maximize the overall association score E_c for limb type c and every subset of allowed connection Z_c (i.e. anatomically plausible connections):

$$\max_{Z_c} E_c = \max_{Z_c} \sum_{m \in D_{j_1}} \sum_{n \in D_{j_2}} (E(d_{j_1}^m, d_{j_2}^n) + \alpha E(\hat{d}_{j_1}^m, \hat{d}_{j_2}^n)) \cdot z_{j_1 j_2}^{mn}, \tag{6}$$

subject to $\sum_{n \in D_{j_2}} z_{j_1 j_2}^{mn} \leq 1, \forall m \in D_{j_1}$ and $\sum_{n \in D_{j_2}} z_{j_1 j_2}^{mn} \leq 1, \forall m \in D_{j_1}$ where

$$\hat{d}_{j_1}^m = \arg\max_{\hat{d}_{j_1}^b} G(d_{j_1}^m, \hat{d}_{j_1}^b), \qquad \hat{d}_{j_2}^n = \arg\max_{\hat{d}_{j_2}^q} G(d_{j_2}^n, \hat{d}_{j_2}^q) \tag{7}$$

are the joints at frame $t-1$ that maximize the temporal consistency along the TAF where b and q span the indexes of the people detected at the previous frame.
In principle, Equation (6) mixes both the contribution coming from the PAF in the current frame and the contribution coming from the PAF obtained by warping, in the previous frame, the candidate joints along the best TAF lines.
In order to speed up the computation, we maximize iteratively Equation (6) by considering only the subsets of joints inside a radius at twice the size of the skeletons in the previous frame at the same location. The complete skeletons are then built, by maximizing, for the limbs type set C, $E = \sum_{c=1}^{C} \max_{Z_c} E_c$.

5 Experiments

We conducted experiments in two different contexts, either on our virtual world dataset JTA and on real data. In the virtual world scenario, we evaluated the capability of the proposed architecture of both reliably extracting people joints and successfully associating them along the temporal dimension. Real data experiments instead, aimed at empirically demonstrating that our virtual world dataset can function as a good proxy for training deep models and to which extent it is necessary to fine-tune the network on real data. In fact, we purposely conducted the experiments either without retraining the network and testing it out-of-the-box or by fine-tuning the network on real data. Moreover, all the tracking experiments do not explicitly model the target appearance, but visual appearance is only taken into account when extracting TAFs, thus exploited only for very short-term target association (namely tracklet construction).

5.1 Experiments on JTA

We tested our proposal on our virtual world scenario in order to evaluate both the joints extraction accuracy and the tracking capabilities. We started from the pre-trained VGG19 weights as the feature extractor and we trained our model end-to-end allowing features fine-tuning. For the temporal branch we randomly split every sequence into 1 second long clips. Subsequently, we uniformly subsampled every clip obtaining 8 frames that are inputted to the temporal branch. The train was performed by using ADAM optimizer with a learning rate of 10^{-4} and batch size equal to 16. We purposely kept the batch size relatively small because every frame carries a high number of different joints at different scales and locations leading to a reliable average gradient for the task.

Detection experiment We first performed a detection experiment in order to quantify the contribution of the individual branch of our architecture. The detection experiment evaluated the location of people joints and the overall bounding box accuracy in terms of detection metrics. Analogously to [33], we used the PCKh (head-normalized probability of correct keypoint) metric, which considers a body joint to be correctly localized if the predicted location of the joint is within a certain threshold from the true location. Table 2 reports the results in term of mean average precision of joints location and bounding box detection metrics such as precision, recall and F1-score with an intersection over union threshold of 50%. We additionally ablated different branch of our architecture in order to empirically measure the contribution of every individual branch (i.e. the occlusion branch and the temporal branch). By observing the Table we can confirm that the network benefits from the presence of the occlusion estimation branch both in terms of joints location accuracy and detection performances. This is due to two different positive effects given by occluded joints. The first is the chance of estimate/guess the position of a person even if visually strong occluded, the second is about maximizing the presence of body joints that greatly simplifies their clustering into skeletons and consequently the detection metrics results improved, Figure 5.(c). Moreover, the temporal branch strengthens this process by adding short-term temporal consistency to the joints location. In fact, results indicate this boosts the performance leading to a more accurate joints detection in presence of people that overlaps in the scene. The improvement is due to the TAFs contribution that helps to disambiguate the association among body joints on the basis of the target direction, Figure 5.(d). Additionally we compared with [7] that was retrained on JTA and tested at 2 different scales (since the method does not deal with multiple scales), against which we score positively. The architecture in [7] is the same as our *Single Image no occ* model in Table 2, with the only difference that the latter has been trained with distance rescaled versions of heatmaps and PAFs, according to Section 4.3, and it deals with multiple scales without any input rescaling operation.

Tracking Experiment We additionally tested the extent of disentanglement between temporal short-term detection and people tracking by performing a complete tracking experiments on the JTA test set. The experiments have been

Table 2. Detection results on JTA Dataset

	Joints	Detection		
	Mean Average Prec.	Precision	Recall	F1 Score
Single Image no occ	50.9	81.5	64.1	71.6
Single Image + occ	56.3	87.9	71.8	78.4
Complete	**59.3**	**92.1**	**77.4**	**83.9**
[7]	50.1	86.3	55.8	69.5

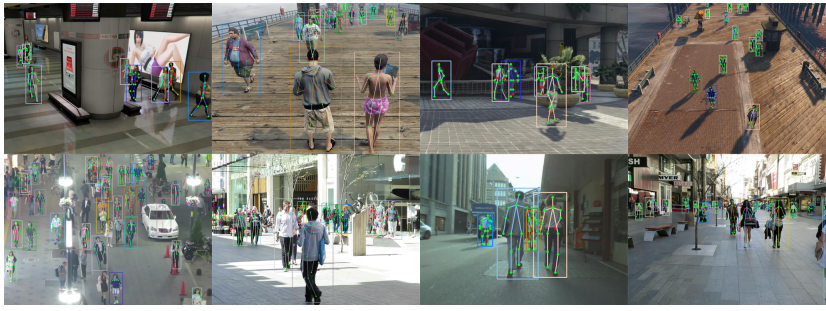

Fig. 6. Qualitative results of THOPA-net on JTA dataset (Top row)and on MOT16 test set(Bottom row)

Table 3. Results on JTA dataset

	MOTA	IDF1	MT	ML	FP	FN	IDs	FRAG
[3] + our det	57.4	57.3	45.3	21.7	40096	103831	15236	15569
[3] + DPM det	31.5	27.6	25.3	41.7	80096	170662	10575	19069
THOPA-net	**59.3**	**63.2**	**48.1**	**19.4**	**40096**	**103662**	**10214**	**15211**

carried out by processing 1 second clips with a stride of 1 frame and associating targets using a local nearest neighbour approach maximizing the TAFs scores. As previously introduced, the purpose of the experiment was to empirically validate the claim that mixing short-term tracking and detection can still provide acceptable overall tracking performance even when adopting a simple association frame-by-frame method. Secondly, this is indeed more evident when the association algorithm exploits more than a single control point (e.g. usually the bounding box lower midpoint), which is the case of tracking sets of joints. For the purpose, we compared against a hungarian based baseline (acting on the lower midpoint of the bounding box), [3], inputed with either our detections and DPM [43], ones. Table 3 reports results in terms of Clear MOT tracking metrics [1]. Results indicate that the network trained on the virtual world scores positively in terms of tracked entities but suffers of a high number of IDs and FRAGS. This behavior is motivated by the absence of a strong appearance model capable of re-associating the targets after long occlusions. Additionally, the motion model is purposely simple suggesting that a batch tracklet association procedure can lead to longer tracks and reduce switches and fragmentations.

Table 4. Results on MOT-16 benchmark ranked by MOTA score

	MOTA	IDF1	MT	ML	FP	FN	IDs	FRAG
[25]	**66.1**	**65.1**	**34.0**	20.8	5061	**55914**	805	3093
[26]	61.4	62.2	32.8	**18.2**	12852	56668	781	2008
THOPA-net	56.0	29.2	25.2	27.9	9182	67059	4064	5557
[27]	47.2	46.3	14.0	41.6	**2681**	92856	774	1675
[28]	46.0	50.0	14.6	43.6	6895	91117	**473**	**1422**
[29]	43.9	45.1	10.7	44.4	6450	95175	676	1795
[30]	38.8	42.4	7.9	49.1	8114	102452	965	1657

5.2 Tracking people in real data

We tested our solution on real data with the purpose of evaluating the generalization capabilities of our model and its effectiveness in real surveillance scenarios. We choose to adopt the MOT-16 Challenge Benchmark [1]. The MOT-16 Challenge consists of 7 sequences in urban areas with varying resolution from 1980 $\times$ 1024 to 640 $\times$ 480 for a total number of approx 5000 frames and 3.5 minutes length. The benchmark exhibits strong challenges in terms of viewpoint changes, from top-mounted surveillance cameras to street level ones, Figure 6. All results are expressed in terms of Clear MOT metrics according to the benchmark protocol [1] and as for the virtual world tracking experiment the tracks were associated by maximizing the TAF scores between detections. A first experiment was conducted on the benchmark training set Figure 7 where we compared the MOTA scored by our model with and without fine-tuning on real data. Fine-tuning was performed by considering the ground truth detections and inserting a default skeleton when our Single Image model scored a false negative obtaining an automatically annotated dataset. The network was subsequently end-to-end fine-tuned, with the exception of the occlusion branch. By observing the results, we can conclude that the features extracted on our virtual world are still capable of extracting people joints in real-world images with a high resolution and sharpness (MOT16-09, MOT16-11) but with limited generalization as the image quality decreases. Nevertheless, even with a limited fine-tuning the network achieves the capability of adapting the features even in presence of a self-annotated dataset with potential errors and inaccuracies. Additionally, Table 4 reports the results of our fine-tuned network compared with the best published state of the art competitors up to now. We include in the Table only online trackers, that are referred on the benchmark website as causal methods [2]. The motivation is that our method performs tracking at low level, using TAFs, for framewise temporal association thus it configures as an online tracker. Additionally, it is always possible to consider our tracklets as an intermediate output and perform a subsequent global association by possibly assessing additional high level information such as strong appearance cues and re-id techniques. Our method performs positively in terms of MOTA placing at the top positions. We observe a high IDS value

[2] Methods that do not perform global batch association across the videos but instead process one frame at a time and share only knowledge about the past.

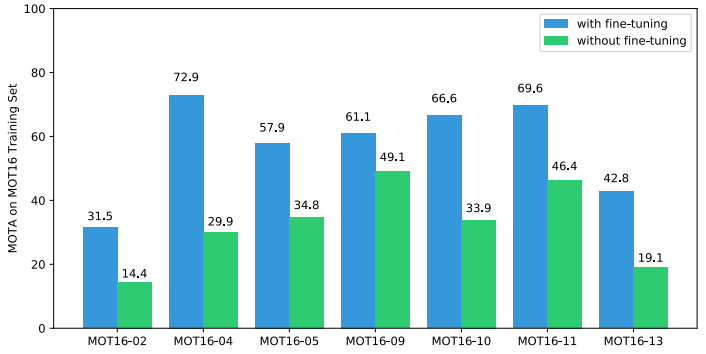

Fig. 7. Bar-chart of THOPA-net (trained on JTA dataset) results on MOT16 training set with and without fine-tuning on real data

and FRAG given by the fact that our output is an intermediate step between detections and long-term tracking. Nevertheless, we remark that we purposely choose a trivial association method that does not force any strong continuity in terms of target trajectories, instead, we argue that given temporal consistency to the target detections the association among them results satisfying for short-term tracking applications. This is possible also thanks to the fact that we use several control points for association (i.e. the joints) that are in fact reliable cues when objects are close each other and the scene is cluttered. Contrary to [25] and [26] our model do not employ strong appearance cues for re-identification. This suggests that the performance can be further improved by plugging a re-id module that connects tracks when targets are lost. Moreover, contrary to [27] we do not employ complex recurrent architecture to encode long-term dynamics. Nevertheless, the performances are comparable suggesting that when a tracker disposes of a plausible target candidate, even if occluded, the association simplify to keep subsequent frames temporally consistent that is indeed what our TAF branch do. Figure 6 shows qualitative results of our proposal.

6 Conclusion

In this paper, we presented a massive CG dataset for human pose estimation and tracking which simulates realistic urban scenarios. The precise annotation of occluded joints provided by our dataset allowed us to extend a state-of-the-art network by handling occluded parts. We further integrate temporal coherency and propose a novel network capable of jointly locate people body parts and associate them across short temporal spans. Results suggest that the network, even if trained solely on synthetic data, adapts to real world scenarios when the image resolution and sharpness are high enough. We believe that the proposed dataset and architecture jointly constitute a starting point for considering

tracking in surveillance as a unique process composed by detection and temporal association and can provide reliable tracklets as the input for batch optimization and re-id techniques. The JTA dataset and code will be publicly released.

References

1. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: MOT16: A benchmark for multi-object tracking. arXiv: 1603.00831 (2016)
2. Andriluka, M., Iqbal, U., Milan, A., Insafutdinov, E., Pishchulin, L., Gall, J., Schiele, B.: Posetrack: A benchmark for human pose estimation and tracking. arXiv preprint arXiv:1710.10000 (2017)
3. Solera, F., Calderara, S., Cucchiara, R.: Towards the evaluation of reproducible robustness in tracking-by-detection. In: 2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). (Aug 2015) 1–6
4. Hamid Rezatofighi, S., Milan, A., Zhang, Z., Shi, Q., Dick, A., Reid, I.: Joint probabilistic data association revisited. In: The IEEE International Conference on Computer Vision (ICCV). (December 2015)
5. Dehghan, A., Tian, Y., Torr, P.H.S., Shah, M.: Target identity-aware network flow for online multiple target tracking. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2015) 1146–1154
6. Feichtenhofer, C., Pinz, A., Zisserman, A.: Detect to track and track to detect. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. (2017) 3057–3065
7. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR. Volume 1. (2017) 7
8. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In Leibe, B., Matas, J., Sebe, N., Welling, M., eds.: Computer Vision – ECCV 2016, Cham, Springer International Publishing (2016) 102–118
9. Gaidon, A., Wang, Q., Cabon, Y., Vig, E.: Virtualworlds as proxy for multi-object tracking analysis. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Volume 00. (June 2016) 4340–4349
10. Carreira, J., Agrawal, P., Fragkiadaki, K., Malik, J.: Human pose estimation with iterative error feedback. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 4733–4742
11. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 4724–4732
12. Hu, P., Ramanan, D.: Bottom-up and top-down reasoning with hierarchical rectified gaussians. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 5600–5609
13. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision, Springer (2016) 483–499
14. Bulat, A., Tzimiropoulos, G.: Human pose estimation via convolutional part heatmap regression. In: European Conference on Computer Vision, Springer (2016) 717–732
15. Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P.V., Schiele, B.: Deepcut: Joint subset partition and labeling for multi person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 4929–4937
16. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In: European Conference on Computer Vision, Springer (2016) 34–50
17. Iqbal, U., Gall, J.: Multi-person pose estimation with local joint-to-person associations. In: European Conference on Computer Vision, Springer (2016) 627–642

18. Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., Murphy, K.: Towards accurate multiperson pose estimation in the wild. arXiv preprint arXiv:1701.01779 **8** (2017)
19. Newell, A., Huang, Z., Deng, J.: Associative embedding: End-to-end learning for joint detection and grouping. In: Advances in Neural Information Processing Systems. (2017) 2274–2284
20. Levinkov, E., Uhrig, J., Tang, S., Omran, M., Insafutdinov, E., Kirillov, A., Rother, C., Brox, T., Schiele, B., Andres, B.: Joint graph decomposition & node labeling: Problem, algorithms, applications. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)
21. Jain, A., Tompson, J., LeCun, Y., Bregler, C.: Modeep: A deep learning framework using motion features for human pose estimation. In: Asian conference on computer vision, Springer (2014) 302–315
22. Zhang, D., Shah, M.: Human pose estimation in videos. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 2012–2020
23. Pfister, T., Charles, J., Zisserman, A.: Flowing convnets for human pose estimation in videos. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 1913–1921
24. Gkioxari, G., Toshev, A., Jaitly, N.: Chained predictions using convolutional neural networks. In: European Conference on Computer Vision, Springer (2016) 728–743
25. Yu, F., Li, W., Li, Q., Liu, Y., Shi, X., Yan, J.: Poi: Multiple object tracking with high performance detection and appearance feature. In Hua, G., Jégou, H., eds.: Computer Vision – ECCV 2016 Workshops, Cham, Springer International Publishing (2016) 36–42
26. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE International Conference on Image Processing (ICIP). (2017) 3645–3649
27. Sadeghian, A., Alahi, A., Savarese, S.: Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In: 2017 IEEE International Conference on Computer Vision (ICCV). (Oct 2017) 300–311
28. Chu, Q., Ouyang, W., Li, H., Wang, X., Liu, B., Yu, N.: Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism. In: 2017 IEEE International Conference on Computer Vision (ICCV). (Oct 2017) 4846–4855
29. Bae, S.H., Yoon, K.J.: Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(3) (March 2018) 595–610
30. Sanchez-Matilla, R., Poiesi, F., Cavallaro, A.: Online multi-target tracking with strong and weak detections. In Hua, G., Jégou, H., eds.: Computer Vision – ECCV 2016 Workshops, Cham, Springer International Publishing (2016) 84–99
31. Andriluka, M., Roth, S., Schiele, B.: People-tracking-by-detection and people-detection-by-tracking. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE (2008) 1–8
32. Izadinia, H., Saleemi, I., Li, W., Shah, M.: 2t: Multiple people multiple parts tracker. In: European Conference on Computer Vision, Springer (2012) 100–114
33. Iqbal, U., Milan, A., Gall, J.: Posetrack: Joint multi-person pose estimation and tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Volume 1. (2017)
34. Insafutdinov, E., Andriluka, M., Pishchulin, L., Tang, S., Levinkov, E., Andres, B., Schiele, B.: Arttrack: Articulated multi-person tracking in the wild. In: IEEE

Conference on Computer Vision and Pattern Recognition (CVPR). Volume 4327. (2017)
35. Zhang, W., Zhu, M., Derpanis, K.G.: From actemes to action: A strongly-supervised representation for detailed action understanding. In: Proceedings of the IEEE International Conference on Computer Vision. (2013) 2248–2255
36. Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J.: Towards understanding action recognition. In: Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE (2013) 3192–3199
37. Charles, J., Pfister, T., Magee, D., Hogg, D., Zisserman, A.: Personalizing human video pose estimation. In: Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on, IEEE (2016) 3063–3072
38. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: Mot16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016)
39. Sapp, B., Weiss, D., Taskar, B.: Parsing human motion with stretchable models. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 1281–1288
40. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. CVPR (2014)
41. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
42. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision, Springer (2014) 740–755
43. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(9) (Sept 2010) 1627–1645

Supplementary Material

Matteo Fabbri Fabio Lanzi Simone Calderara Andrea Palazzi
Roberto Vezzani Rita Cucchiara

University of Modena and Reggio Emilia, Italy

1 Per Joint Results on JTA

Table 1 reports the results in term of mean average precision (mAP) per joints.

Table 1. Mean Average Precision (mAP) per body joint. Experiments are performed on JTA Dataset

	Head	Shou	Elb	Wri	Hip	Knee	Ankl	Total
Single Image no occ	63.5	55.1	48.4	41.1	55.0	46.4	38.5	50.9
Single Image + occ	70.5	61.3	53.7	45.9	61.0	51.6	42.8	56.3
Complete	**74.4**	**64.8**	**56.9**	**48.5**	**64.3**	**54.5**	**45.1**	**59.3**
Cao *et al.*	62.5	54.3	47.5	40.6	54.0	45.5	37.9	50.1

2 Per Sequence Results on MOT-16

Table 2 reports the metrics per sequence performed on MOT-16.

Table 2. Results on MOT-16 benchmark per sequence

Sequence	MOTA	IDF1	MT	ML	FP	FN	IDs	FRAG
MOT16-01	36.8	30.8	30.4	13.0	1110	2710	222	280
MOT16-03	71.6	34.7	46.6	10.1	1156	26839	1723	2454
MOT16-06	55.1	14.4	31.7	29.0	721	4159	302	323
MOT16-07	41.9	29.8	18.5	16.7	1233	7759	489	713
MOT16-08	32.5	22.7	15.9	34.9	862	10109	327	476
MOT16-12	38.5	20.7	20.9	38.4	958	4027	115	247
MOT16-14	16.2	13.0	4.3	40.2	3142	11456	886	1064

3 JTA Dataset

Figure 1 and Figure 2 provide examples from JTA dataset exhibiting its variety in viewpoints, number of people, illuminations and scenarios.

Fig. 1. Some images taken from JTA dataset

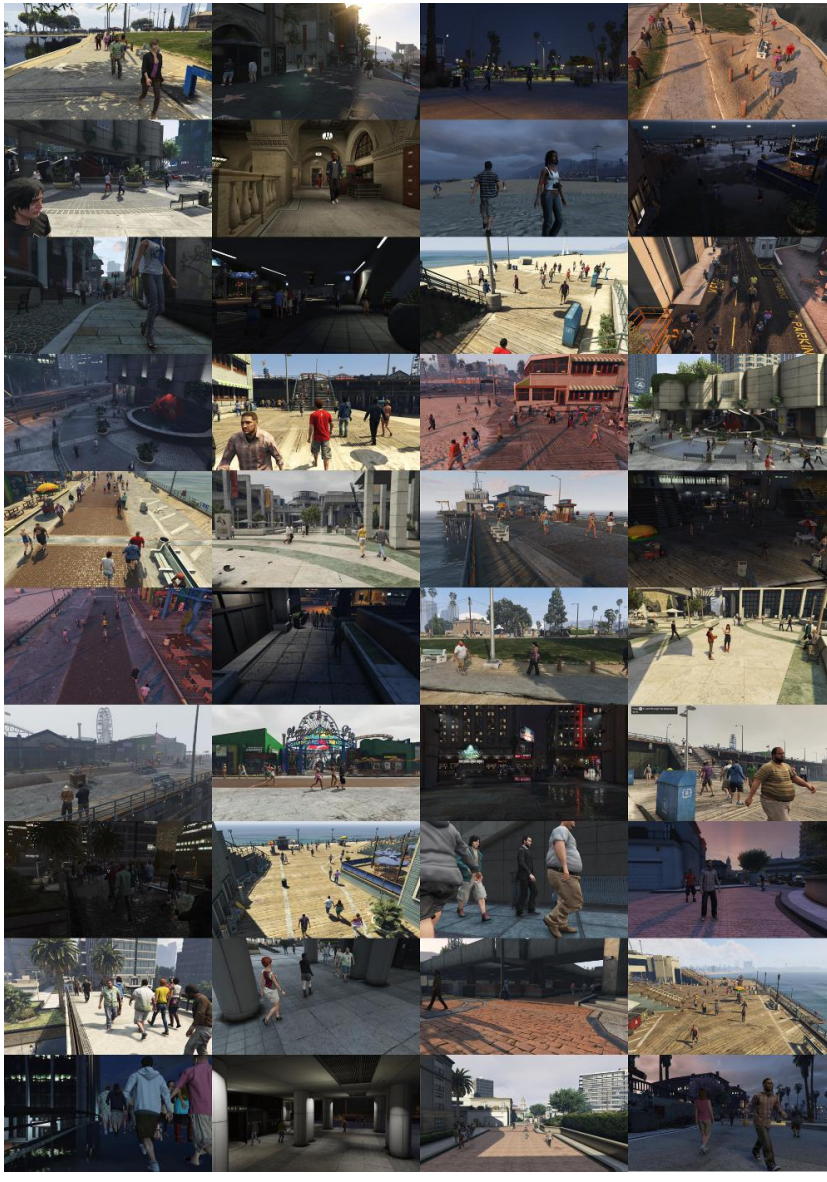

Fig. 2. Some images taken from JTA dataset