Fine-Grained Human Analysis Under Occlusions and Perspective Constraints in Multimedia Surveillance

RITA CUCCHIARA and MATTEO FABBRI, Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia, Italy

Human detection in the wild is a research topic of paramount importance in Computer Vision and it is the starting step for designing intelligent systems oriented to human interaction that works in complete autonomy. To achieve this goal, Computer Vision and Machine Learning should aim at superhuman capabilities. In this work, we address the problem of fine-grained human analysis under occlusions and perspective constraints. More specifically, we discuss some issues and some possible solutions to effectively detecting people using pose estimation methods, to detect humans under occlusions both in the 2D image plane and in the 3D space exploiting single monocular cameras. Dealing with occlusion can be done at joint level or pixel-level: we discuss two different solutions, the former based on a supervised neural network architecture for detecting occluded joints and the former based on a semi-supervised specialized GAN which exploits both appearance and human shape attributes, to hallucinate the missing parts of the visible shape. To deal with perspective constraints, we further discuss a neural approach based on a double-architecture that learns to create an optimal neural representation, useful to reconstruct the 3D position of human keypoints starting with simple RGB images. All these approaches have a critical point in common, that is the need for large annotated datasets. To have large, fair, consistent, transparent, and ethic-complaint datasets we propose the adoption of synthetic datasets as, for example, JTA and MOTSynth. In this paper, we discuss the pros and cons of using synthetic datasets while tackling several human-centered AI issues in respect of European GDPR rules for privacy. We further explore and discuss an application in the field of risk assessment by space occupancy estimation during the COVID-19 pandemic, called Inter-Homines.

CCS Concepts: • **Computing methodologies** → **Tracking**; **Object detection**; **Reconstruction**.

Additional Key Words and Phrases: People Detection, Human Pose Estimation, Tracking, 3D Localization, Synthetic Dataset

ACM Reference Format:
Rita Cucchiara and Matteo Fabbri. 2021. Fine-Grained Human Analysis Under Occlusions and Perspective Constraints in Multimedia Surveillance. *ACM Trans. Multimedia Comput. Commun. Appl.* 1, 1, Article 1 (January 2021), 22 pages. https://doi.org/10.1145/3476839

1 INTRODUCTION

Video-Surveillance is one of the most classical applications for Computer Vision and Multimedia. In particular, Multimedia Surveillance has been deeply explored in the last fifteen years. Multimedia surveillance refers to the study and understanding of real scenes for interactive tasks, where the automatic processes of identifying phenomena of interest and forecasting anomalous situations are designed to empower human capabilities in monitoring for safety and security [8]. The targets could be whatever moving object that could appear in the scene: people, vehicles, animals, or even weapons.

Authors' address: Rita Cucchiara, rita.cucchiara@unimore.it; Matteo Fabbri, matteo.fabbri@unimore.it, Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia, Via P. Vivarelli 10, Modena, Italy, 41125.

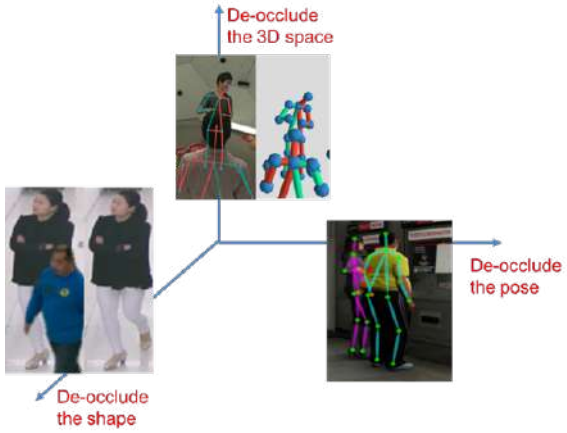

Fig. 1. Three dimensions for de-occluding fine grained human detections.

Indeed, humans and their anomalous activities are the main focus of recent research in surveillance: human detection, human motion tracking, people re-identification, and human behavior understanding are some of the tasks that contribute to finding solutions in Multimedia Surveillance for human users. Let's consider, for instance, systems that analyze the risk of COVID contagion: those systems should be conceived not only as mere automatic social distancing systems, capable of measuring the mutual distance of people to keep them far enough from each other, but they should be designed to give multimedia information about the risk of contagion in an area by modeling the space occupancy and forecasting risk information for human controllers. To this aim, we need a new generation of surveillance systems capable of seeing better than their human counterpart, despite relying on a much simpler monocular vision.

Video processing surveillance has been scientifically explored for more than 20 years, starting from the famous Pfinder MIT program [49], followed by several other projects in the early twenties [9, 19]. Even though many products for automatic surveillance are now installed everywhere, the problem of having a complete human behavior understanding is still far from being solved.

In this paper, we discuss a very specific aspect, namely the ability of deep learning-based architectures to see humans and their fine-grained details with superhuman vision, i.e., with a technology that could be better than humans by design. In particular, this paper proposes some discussion points and shows some possible solutions for the problem of detecting humans in extreme conditions in surveillance contexts: estimating their aspect and their pose under severe occlusions and estimate their 3D position under unknown perspective constraints due to the distance between target and camera. These challenges are difficult and mostly impossible for human sight but are strategic for surveillance and multimedia applications for human behavior understanding.

The people detection problem, which is the starting step for all human-centered surveillance tasks, ideally is not complex per-se, since the human shape presents low variability and a very characteristic boundary aspect. However, "people detection in the wild" is still an open problem as there is not a single solution capable of detecting people from whichever view and for partial or occluded views, as well as to detect their position in the 3D world space, employing a simple monocular camera.

For many years the main constraint has been the use of fixed calibrated cameras, to have the required intrinsic and extrinsic camera parameters to reconstruct the geometry of the scene. This constraint has recently been relaxed thanks to supervised learning approaches. In particular, artificial neural networks are now able to produce accurate 3D locations for an undefined number of people by solely relying on an RGB image [15, 28, 30]. Those solutions can be easily employed for processing videos from PTZ and moving cameras (also mounted on vehicles or moving robots) as well as for elaborating multimedia data taken from the web.

Recently, people detection approaches inputting image frames and outputting bounding boxes as human descriptors also exploit deep learning architectures. Among them, YOLOv3 [32], CenterNet [56] and Faster R-CNN [33] are the most widely utilized convolutional neural networks. Those approaches have been coupled with other methods based on pose estimation [37, 50]. Technically, pose estimation refers to the capability of localizing the main body joints of a person [6, 16]. Pose estimation approaches are generally adopted for fine-grained people analysis but they often replace regular people detectors as they are usually more robust to occlusions.

This work does not intend to be a review of the methods that target human analysis in surveillance, but instead, it aims at discussing some possibilities of going beyond human vision with artificial systems. How can we design artificial systems capable of having a fine-grained understanding of people that is superior to human capabilities? Here, we do not consider augmented sensors such as high-resolution, high-frame cameras, thermal, depth, stereo, or event cameras, but we rather address the vision-related problems from a single monocular camera.

In order to provide humans with tools that could be useful empowering instruments to enrich their capabilities, we analyze the problem of de-occluding people. Thus, a more specific question could be: how can we achieve a fine-grained understanding of people even under severe visual occlusions or perspective aberrations? The answer can be formulated by looking at a three-dimensional space of search, as in Figure 1, where the three dimensions refer to i) how to detect people and their fine-grained poses under severe occlusions, ii) how to reconstruct the person aspect and people shape in presence of missing information, iii) how to reconstruct occluded people position in the 3D space from a monocular camera.

For each of these three lines of research, we present some recent solutions proposed at AImageLab UNIMORE, Italy, with a special focus on critical discussion on results and limitations. In particular, we will describe i) a neural architecture that focuses on finding occluded body joints in order to discriminate and better detect overlapping people in surveillance environments, ii) a semi-supervised approach based on a triple-discriminative Generative Adversarial Network tasked to fill the missing parts of occluded people and iii) a bottom-up approach based on an auto-encoder which learned how to compress people pose representations in the space. All these methods have in common a powerful paradigm: learning by synthetic data in virtual environments, which will be discussed as well.

2 LEARNING HUMANS FROM SIMULATED DATA

Modern Machine Learning has a basic statement i.e., neural networks require a huge amount of training data, and, especially for current supervised or semi-supervised approaches, data is never enough: the more the better. But we could also say that this is not true in general as we also need “good” data with a large variety and uniformly distributed redundancy. Moreover, in order to create correct and transparent AI solutions, datasets should be collected with fairness.

As data concern humans, the first critical issue is the type of collected human data. The data variety should be respectful of all human-related issues regarding privacy, gender balance, and other important values as defined in the “White Paper on Artificial Intelligence” [1], depending on the scope of the data collection and data processing. In this paper, we do not address the problem of human identity and neither any other issue that could affect ethics, since the

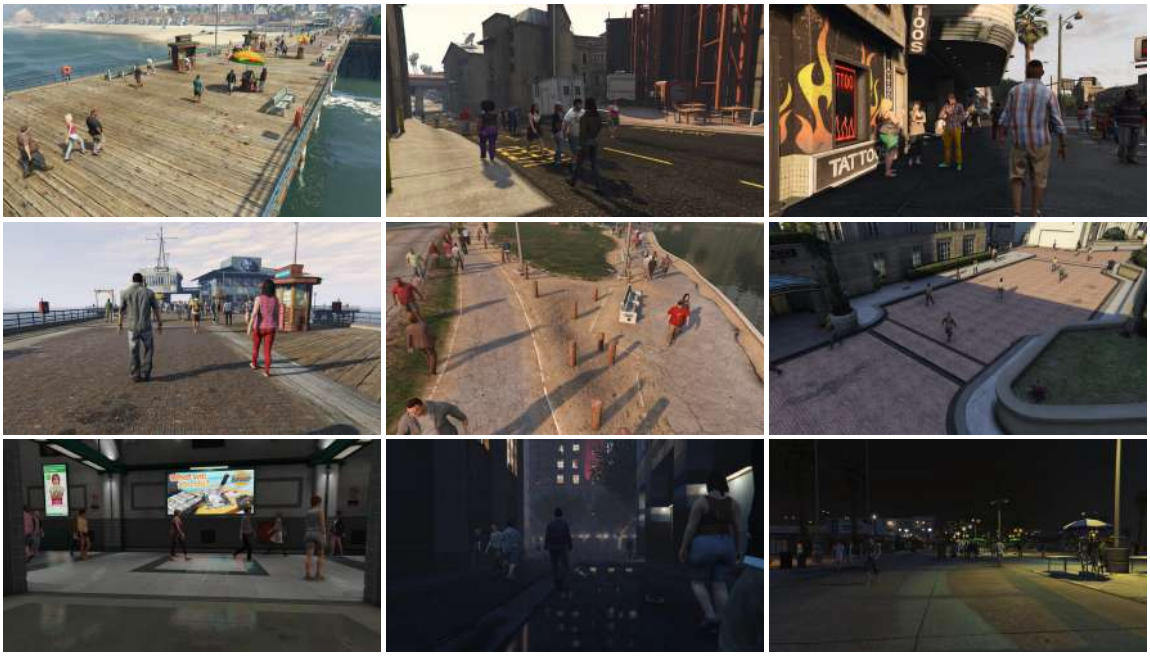

Fig. 2. Examples from the JTA dataset exhibiting its variety in viewpoints, number of people and scenarios. @Rockstar Games, Inc.

task is to detect people and their position without storing any sensitive information. Although we know that each technique can have a dual-use for unethical purposes, we aim at designing valuable applications for our society, like security systems to recognize anomalous activities (e.g. vandalism and shoplifting); techniques for safety purposes such as monitoring systems for the contagion risk in crowded areas or vision products that assess the safety of workers in machine-robot interaction; systems designed for statistical analysis with the goal of economic sustainability or systems applicable to sport surveillance. For all these applications, a tremendous amount of work is put on annotating a massive quantity of data with information concerning, ad example, the people position in the three-dimensional space, their posture, or their attributes useful for statistical purposes (e.g. gender, age, wearing glasses or carrying a pack).

Another critical issue regards the distribution of diversity in the data collection. The data must be well representative of the elements that we would like to describe: humans in our case. We must ensure that a trained algorithm capable of detecting people is independent of the human race, gender, age, dressing style, and other appearance properties to be accurate and exhaustive. Many datasets have been proposed for people detection, people pose estimation and tracking such as MOT-17 [29], MOT-20 [11] and PoseTrack [2]. The data acquisition and annotation of those datasets required an enormous amount of manual effort. Indeed, manual annotation inherits all the drawbacks connected with the limited capabilities of the human senses. Manual annotation can be:

- error-prone due to human errors (e.g., missing annotations);
- imprecise because of human inaccuracy (e.g., bounding boxes can be too tight or too loose for object detection);
- unaffordable due to the annotation cost (e.g., instance segmentation in videos with hundreds of people at 30 fps);
- inconsistent for subjective tasks (e.g., determining the age of a person in attribute recognition);
- unfeasible due to the need for different sensors (e.g., 3D pose estimation in public crowded areas);

- impossible because of missing information (e.g., annotation of occluded body joints for 2D pose estimation).

As more data is constantly required to train ever-growing models, the effort required for collecting such datasets is becoming prohibitive. This burden can either limit the quantity or the quality of data acquired, slowing down the progress in Computer Vision. If we want to reach superhuman capabilities in AI we should put a special effort into collecting "superhuman" datasets, thanks to which AI solutions can learn superhuman abilities. A possible way of providing superhuman datasets while also providing solutions to the aforementioned problems is to employ virtual worlds.

An example of a synthetic dataset for human behavior understanding is the Joint Track Auto (JTA) dataset [16] produced at AImageLab - UNIMORE and collected using the highly photorealistic *Grand Theft Auto V* videogame developed by *Rockstar North*. JTA has been conceived for making automatic annotation available for the community, in order to speed up the research in many Computer Vision fields. The annotations provided encompass many tasks like people detection, people tracking and multi-person 2D and 3D pose estimation. In Figure 2 some examples of JTA are presented.

The first version of the JTA dataset contains 512 clips recorded for surveillance purposes. The collected videos feature a vast number of different body poses, in several urban scenarios at varying illumination conditions and viewpoints. The dataset also contains moving sequences where the camera moves through the crowd. It contains almost half a million frames and about 10 million poses with a range of 0 to 60 people per frame. The majority of people walk or stay in a still position but it is sometimes possible to spot people sitting on a bench or running. People's gender and ethnicity are balanced. Every clip provides a precise annotation of visible and occluded body parts, as well as people tracking with 2D and 3D key-point locations in the standard camera system. JTA overcomes most of the limitations of existing datasets in terms of volume of data.

Data acquisition has been carried out using a tool that allows the integration of native functions of the video game in custom scripts. Those scripts are generally used by players to create game modifications (mods) that alter one or more aspects of the video game, such as how it looks or behaves. For the creation of JTA, we took advantage of the full

Table 1. Comparison on MOT17 against synthetic and real datasets.

	Dataset	AP	MODA	FAF	TP	FP	FN	Rec.	Pr.
YOLOv3	COCO [25]	69.76	62.02	1.25	47824	6650	18569	72.03	87.79
	VIPER [35]	26.65	22.02	0.16	15447	838	50910	23.28	94.85
	JTA [16]	53.18	48.77	0.79	36578	4200	29815	55.09	89.70
	MOTSynth–256	62.99	62.31	0.58	44458	3090	21935	66.96	93.50
	MOTSynth	71.90	64.51	1.07	48500	5673	17893	73.05	89.53
CenterNet	COCO [25]	67.01	44.38	3.37	47398	17935	18995	71.39	72.55
	VIPER [35]	44.58	36.92	1.24	31122	6611	35271	46.88	82.48
	JTA [16]	60.15	45.38	2.32	42435	12308	23958	63.91	77.52
	MOTSynth–256	61.82	50.11	2.03	44067	10795	22326	66.37	80.32
	MOTSynth	70.49	55.25	2.11	47883	11204	18510	72.12	81.04
FR-CNN	COCO [25]	76.68	53.86	3.45	54127	18364	12266	81.52	74.67
	VIPER [35]	60.93	42.87	2.87	43707	15241	10593	65.82	74.14
	JTA [16]	69.69	38.38	5.12	52726	27242	13667	65.93	79.41
	MOTSynth–256	78.61	58.65	3.10	55441	16504	10952	83.50	77.06
	MOTSynth	78.98	54.96	3.51	55121	18634	11272	83.02	74.74

Table 2. Overview of the publicly available datasets for pedestrian detection and tracking. For each dataset, we report the numbers of clips, annotated frames and instances. We also report the presence of 3D data and occlusion information, as well as the availability of labels for pose estimation, instance segmentation, and depth estimation. The next column shows the data type: autonomous driving (AD), diverse (DV) or urban surveillance (US). Last two columns provide information about publication conference or journal and year of publication.

Dataset	#Clips	#Frames	#Instances	3D	Occl.	Pose	Segm.	Depth	Type	Pubbl.	Year
KITTI [18]	50	22k	160k	✓	✓			✓	AD	CVPR	2012
nuSCENES [5]	1,000	40k	280k	✓					AD	CVPR	2020
BDD100k-MOTS [52]	70	14k	129k		✓		✓		AD	TDV	2018
BDD100k-MOT [52]	1,600	100k	3,300k		✓				AD	TDV	2018
Waymo Open [43]	1,150	230k	2,700k	✓					AD	CVPR	2020
PoseTrack [2]	1,356	46k	276k			✓			DV	CVPR	2018
MOTS [45]	4	3k	27k		✓		✓		US	CVPR	2019
MOT-17 [29]	14	11k	293k		✓				US	arXiv	2016
MOT-20 [11]	8	13k	1,652k		✓				US	arXiv	2020
VIPER [35]	187	254k	2,750k	✓	✓		✓		AD	ICCV	2017
GTA [23]	-	250k	3,875k				✓	✓	DV	CVPR	2018
JTA [16]	512	460k	15,341k	✓	✓	✓			US	ECCV	2018
MOTSynth [13]	768	1,382k	40,781k	✓	✓	✓	✓	✓	US	ICCV	2021

potential of the videogame by altering the weather, the time of day, the camera position, and the people's appearances and behaviors. Specifically, we utilized two different mods: one for the scenario creation and one for the actual recording. Using a film-making analogy, the first mod represents the pre-production where the screenplay is written and the various locations are chosen while the second mod consists in the actual production stage where raw footage and other elements are recorded. A straightforward advantage of using synthetic datasets is that we can annotate invisible details of humans, that is, for instance, the occluded people joints, thus providing a superhuman visual annotation.

An open question remains: how much synthetic datasets are useful when a network trained on them is employed in real-world scenarios? Some discussions can be found in Fabbri et al. [16]. The answer is the same as we were discussing the usefulness of real but limited datasets. The generalization capabilities of networks trained on a dataset are constrained by the variety and by the completeness of the dataset itself. In our first paper [16], we observed that results are good after a small fine-tuning that copes with domain-shift-related problems. In general, training solely on JTA and testing on real-world scenarios do not yield good performance due to the low diversity of pedestrian appearance and low variability of camera position.

In order to better understand the problems related to the domain shift between synthetic and real data, we recorded a second version of JTA: MOTSynth [13]. The improved version has three times the number of annotated frames with a higher variety of environments, camera position, and pedestrian models. Moreover, along with 2D and 3D pose annotations, the new dataset also provides ground-truth for instance segmentation, and depth estimation. All the almost 1.4 billion frames are densely annotated at 25 fps. Global IDs are also provided for re-identification purposes. Preliminary results leveraging the newly recorded dataset show superior performances when compared against real-world datasets like COCO [25].

To understand how training on MOTSynth compares to large-scale real-world datasets, we perform a series of experiments involving three heterogeneous object detectors: Faster RCNN [34] as two-stage detector, YOLOv3 [32] and

Fig. 3. Examples from the AiC dataset exhibiting its variety in viewpoints, illuminations and scenarios. @Rockstar Games, Inc.

Table 3. Overview of the publicly available datasets for Human Attribute Classification. For each dataset we reported the numbers of scenes, the number of samples, as well as the number of annotated visual attributes, the image resolution, publication journal or conference, and year of publication.

Dataset	# Scenes	# Samples	# Attributes	Resolution	Publication	Year
PETA [12]	-	19,000	61(+4)	17×39 to 169×365	ACMM	2014
Market-1501 [55]	-	34,213	13	63×128	ICCV	2015
RAP [24]	26	41,585	69(+3)	36×92 to 344×554	arXiv	2016
PA-100K [26]	598	100,000	26	50×100 to 758×454	ICCV	2017
AiC [17]	512	125,000	24	35×85 to 602×1080	CVIU	2019

CenterNet [56] as single-stage detectors. For each detector, we compared MOTSynth training against COCO training by testing on MOTChallenge.

As shown in Table 1, MOTSynth training clearly outperforms the real COCO dataset and alternative synthetic datasets consistently. What is the advantage of MOTSynth over MOTSynth – is it the diversity or sheer amount of data? To answer this question, we conduct the following experiment. We train each detector using the subset of MOTSynth, MOTSynth–256, containing only 256 sequences, generated from the screenplays used to generate [16]. The only difference between JTA and MOTSynth–256 is in people appearance variation – high person appearance variety was one of the key goals when generating MOTSynth sequences. As can be seen, with YOLOv3 and Faster R-CNN MOTSynth–256 models, we obtain +9.81*AP* and +8.92*AP* over JTA trained models. This shows that the MOTSynth diversity in terms of people appearance is a crucial ingredient for bridging the domain gap.

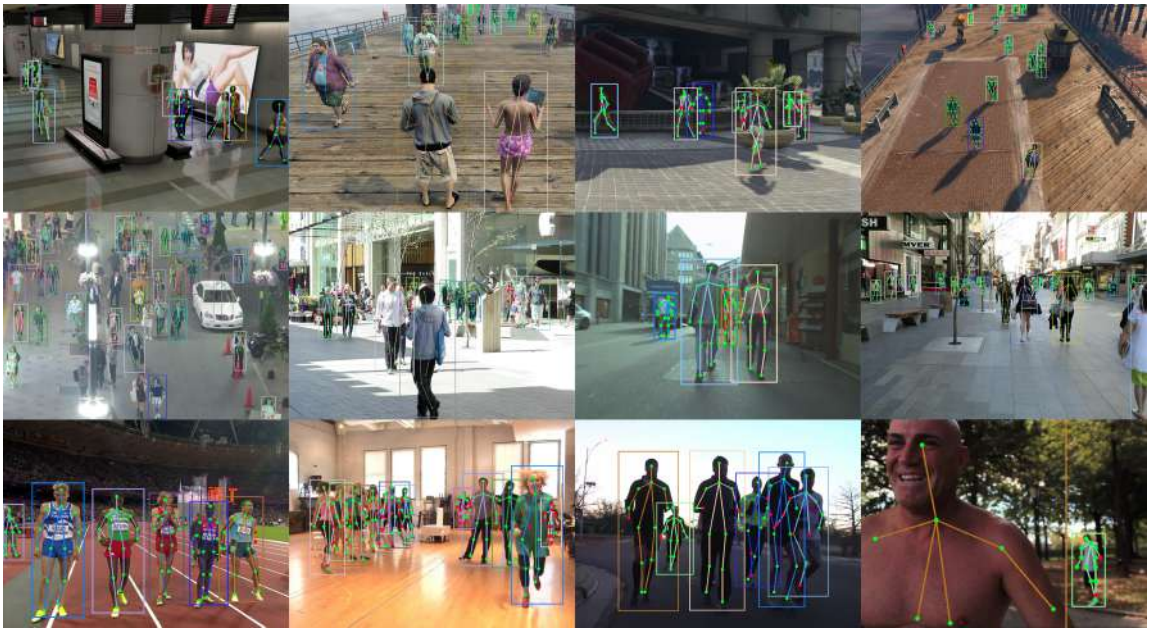

Fig. 4. Results in a real setting. The Figure is taken from the paper "Learning to Detect and Track Visible and Occluded Body Joints in a Virtual World" [16]

Table 2 shows the most widely used publicly available datasets for people detection and people tracking in videos. Both JTA and MOTSynth, which focus on urban scenarios, are superior in terms of the number of frames and types of annotations. In particular, MOTSynth contains two orders of magnitude more frames than manually annotated datasets like PoseTrack and MOT-17, while having a richer annotation that encompasses 3D keypoint location, occlusion information, instance segmentation, and depth data.

Similar considerations can be done for existing datasets dealing with human attribute classification. Most of the publicly available pedestrian attribute datasets, like RAP [24], Market-1501 [55], PETA [12] and PA-100K [26] does not contemplate occlusion events. They only provide samples of fully visible people, completely ignoring crowded situations of pedestrians occluding each other (which is indeed common in urban scenarios). To overcome this limitation, we collected Attributes in Crowd dataset [17], a synthetic dataset for people attribute recognition in presence of strong occlusions. AiC features 125,000 samples, all being unique subject, each of which is automatically labeled with information concerning sex, age, etc. Each of the 24 attributes occurs at least in 10% of samples which highlights a good balance in terms of labels. Each image sample has its vanilla version where each obstacle is removed from the image. Thus, for each occluded pedestrian, we know exactly how it really is behind the occlusion (this is indeed not achievable in real environments). Fig. 3 exhibits some examples of AiC while Table 3 shows the comparison against other publicly available datasets.

3 HUMAN DETECTION BY POSE COPYING WITH OCCLUSIONS

A first dimension for supporting humans with artificial detection systems is to provide missing pose information, that is, the estimation of the position of occluded or self-occluded body joints. The value of such solutions is straightforward: it

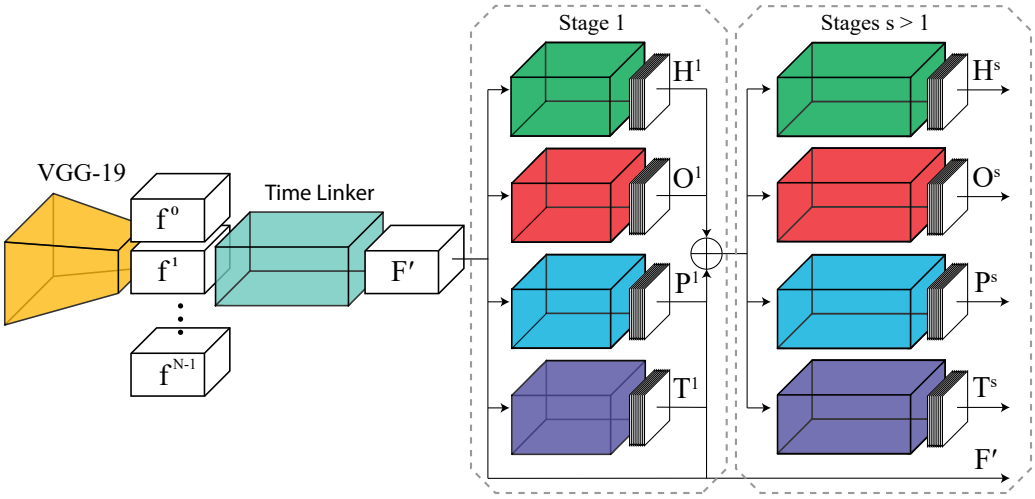

Fig. 5. The THOPA-Net architecture for occlusion pose detection and tracking. The VGG-19 backbone takes N frames as input and produces N intermediate representations $f^0, f^1, ..., f^{N-1}$. The N representations are fed to the Time Linker to produce a single set of feature maps F' which are subsequently processed by a three-branch multi-stage CNN where each branch focuses on a different aspect of body pose estimation: the first branch predicts the heatmaps H of the visible parts, the second branch predicts the heatmaps O of the occluded parts, the third branch predicts the part affinity fields P, which are vector fields used to link parts together in space, and the fourth branch predicts the temporal affinity fields T that links parts together in time.

could be useful to avoid false negatives or to have a more precise detection that makes tracking solutions more robust in crowded scenarios where people often occlude each other. As well, in the case of not overlapping people, understanding occluded joints could be useful to estimate the motion, the direction, and the people activity.

A simple but effective method to detect occluded body joints is THOPA-Net (Temporal Heatmaps and Occlusions based body Part Association) [16] which improves the architecture in [6] by taking into account the occlusion and the motion of every joint in the image.

THOPA-net jointly extracts people's body parts and associates them across short temporal spans. The model explicitly deals with occluded body parts, by hallucinating plausible solutions of not visible joints. The architecture trained on JTA 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. Indeed, temporal continuity in the detection phase gains more importance when scene cluttering introduces the challenging problems of occluded targets. Figure 4 shows some qualitative results of the method.

More specifically, the 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, it integrates a branch for handling occluded joints in the detection process. Subsequently, a temporal linking network integrates temporal consistency by jointly achieving detection and short-term tracking. The Single Image model takes an RGB frame as input and produces, as output, the pose prediction for every person in the image. Conversely, the complete architecture (Figure 5) takes a clip of N frames as input (e.g. $N = 8$) and outputs the pose prediction for the last frame of the clip and the temporal links with the previous frame.

Table 4. Tracking Results on JTA Dataset

	MOTA	IDF1	MT	ML	FP	FN	IDs	FRAG
Solera *et al.* [42] + our det	57.4	57.3	45.3	21.7	40096	103831	15236	15569
Solera *et al.* [42] + 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**

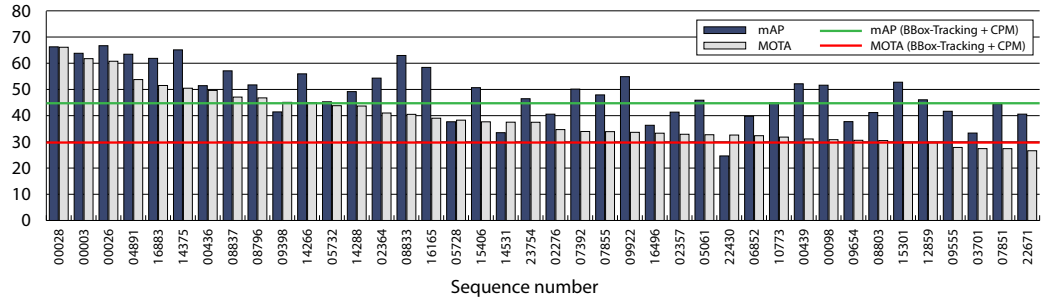

Fig. 6. Results on PoseTrack dataset compared with a BBox-Tracking + CPM (trained on MPII) baseline (used also in [20]; red/green lines are the average of performances on the selected sequences to avoid plot clutter)

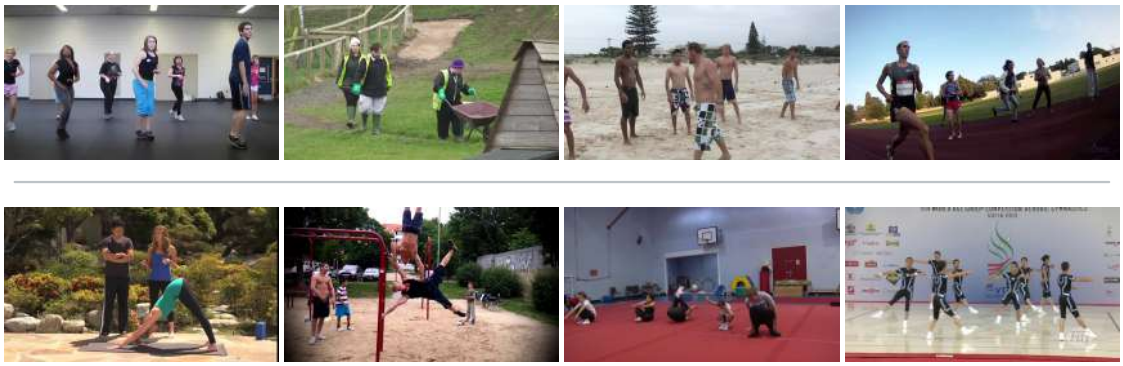

Fig. 7. Samples taken from Posetrack. First row: sequences with low pose variability (sequence numbers 00028, 00003, 00026, and 04891). Second row: sequences with high pose variability (sequence numbers 003223, 007128, 009268, and 009521).

This supervised approach was only possible thanks to synthetic data for two reasons: the first is that virtual data is always complete and precise. In fact, 3D coordinates are always available even when the person is under complete occlusion. The second reason is that, for precise localization, the target's camera distance information should be exploited during training. In fact, people far from the camera look way different than people close to the camera. For this reason, they should be differentiated during training in order to help the network have a richer understanding of the world. Specifically, 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 $p \in \mathbb{R}^2$ is given by:

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

Table 5. Results on MOT-16 benchmark ranked by MOTA score

	MOTA	IDF1	MT	ML	FP	FN	IDs	FRAG
Yu *et al.* [51]	**66.1**	**65.1**	**34.0**	20.8	5061	**55914**	805	3093
Wojke *et al.* [48]	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
Sadeghian *et al.* [39]	47.2	46.3	14.0	41.6	**2681**	92856	774	1675
Chu *et al.* [7]	46.0	50.0	14.6	43.6	6895	91117	**473**	**1422**
Bae *et al.* [4]	43.9	45.1	10.7	44.4	6450	95175	676	1795
Cavallaro *et al.* [40]	38.8	42.4	7.9	49.1	8114	102452	965	1657

where σ regulates the spread of the peak in function of the distance d of each joint from the camera.

Having precise joints positions, of both visible and occluded ones, is essential to disambiguate people in-crowd. In fact, in our work, we used the joint positions, the PAFs (Part Affinity Fields), and the TAFs (Temporal Affinity Fields) to assess the spatiotemporal coherency of each person with high accuracy and displaying superhuman capabilities.

Table 4 reports results in terms of Clear MOT tracking metrics [29] obtained on JTA. Results indicate that the network trained on the virtual world scores positively in terms of tracked entities but suffers from 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.

The main limitation, as previously stated, is on the capability of the network trained solely on synthetic datasets to generalize in real scenes. We tested the network on different real contexts using the Posetrack dataset and we showed that domain adaptation is possible only if people posture and movement are consistent between the two domains. In figure 6 MOTA and mAP per-sequence results of THOPA-net on PoseTrack are shown. The plot only shows the 40 sequences that obtained the best results.

Figure 7 (first row) shows some samples taken from the top four scoring sequences. As can be seen, the postures of the subjects are similar to the ones provided by the training set of JTA, as people are walking or running. Figure 7 (second row), on the other hand, shows some samples collected from the sequences where our method failed to properly predict human poses. In fact, the pose variability of those sequences does not align with the training set of JTA.

In general, results are satisfying even if the network is trained solely on CG data, suggesting it could be a viable solution for fostering research in the joint tracking field, especially for urban scenarios where real joint tracking datasets are missing.

Additionally, we fine-tuned THOPA-Net on MOT-16 training set, with the exception of the occlusion branch. Table 5 reports the results of our fine-tuned network compared with state-of-the-art competitors. We include in the table only online trackers. Our method performs positively in terms of MOTA placing at the top positions, showing that fine-tuning on real data is still required to bridge the gap between synthetic and real domains.

4 HUMAN APPEARANCE HALLUCINATION UNDER OCCLUSIONS

Supervised learning can be adopted for training a network to recognize occluded joints by predicting heatmaps where, for each pixel, a corresponding value indicates the probability that there is an occluded joint in that specific location.

A more complex task is to hallucinate the occluded parts of a body when not visible. This is a relatively simple cognitive exercise for humans that have been constantly trained to see people, their clothes, and their aspect throughout

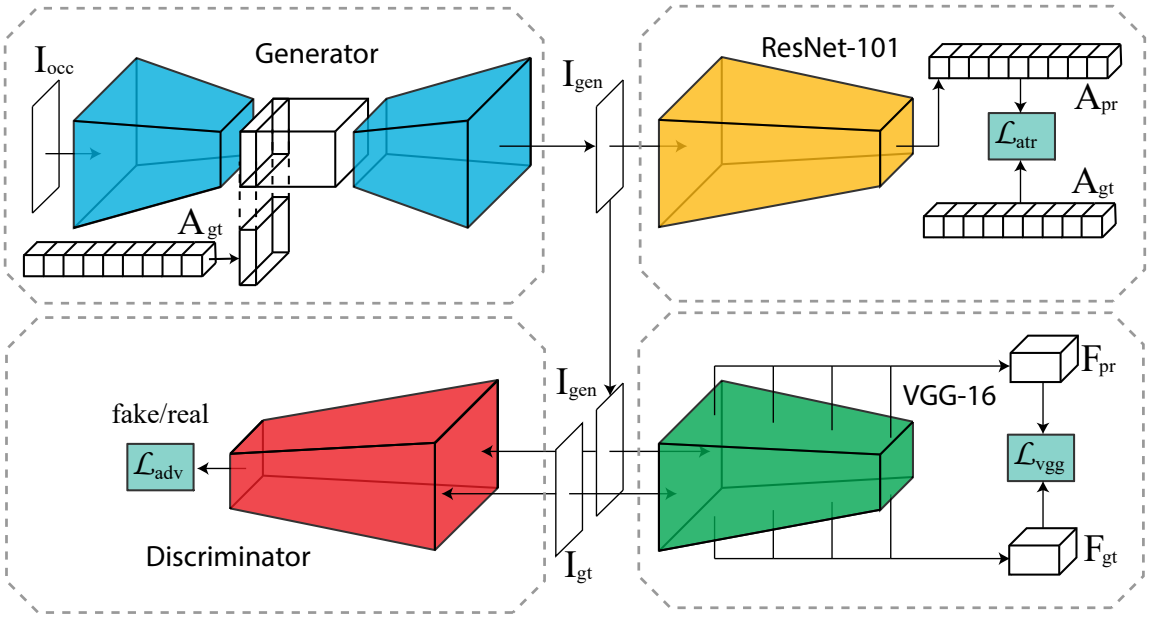

Fig. 8. A schematic representation of the training procedure adopted in our work. The Generator takes the occluded image I_{occ} as input and the attributes of the person A_{gt} as a further conditioning element. To train the Generator, we fed the generated image to three different networks: ResNet-101, VGG-16, and the Discriminator in order to compute the relative losses. ResNet-101 is used to maximize high-level similarity. VGG-16 is used to encourage low-level similarity. The Discriminator, which gives the judgment between "real" and "fake" distributions, has to be fooled by the Generator in order to produce images belonging to the non-occluded domain of pedestrians.

their lives. Probably, if we ask a person to draw the missing part of a body shape we will achieve satisfactory results. But, at the same time, it is unfeasible to create a large and manually annotated training set containing couples of the same instance of occluded and visible people. To this aim, computer graphics come again to our aid with the AiC [17] dataset.

In this context, semi-supervised methods like GANs (Generative Adversarial Nets) are particularly suitable. The basic idea is to train a conditioned-GAN to generate an image of a person that could be virtually acceptable, i.e., that could be precise enough to confuse a discriminator tasked to distinguish between fake and real images. This was a former approach followed in [14] where we proposed a GAN to create a de-occluded version of a person.

In order to improve the results, we enriched the adversarial paradigm where a more complex Generative Adversarial Network has been conditioned on three objectives. Specifically, the reconstructed image is i) without occlusion ii) similar at pixel level to its completely visible version iii) capable to conserve similar visual attributes (e.g. male/female) of the original one. As depicted in Figure 8, the network is trained to optimize a Loss function which takes into account the three aforementioned objectives:

$$\mathcal{L}_{total} = \overbrace{\underbrace{\mathcal{L}_{adv}}_{\text{adver. loss}} + \underbrace{\lambda_1 \cdot \mathcal{L}_{vgg}}_{\text{cont. loss}} + \underbrace{\lambda_2 \cdot \mathcal{L}_{atr}}_{\text{attr. loss}}}^{\text{total loss}} \tag{2}$$

Table 6. Ablation study results on RAP dataset

Method	mean Accuracy	Accuracy	Precision	Recall	F1	SSIM	PSNR
Baseline	70.74	56.55	70.61	71.78	71.19	0.7982	20.31
VGG loss	72.48	58.89	72.58	73.56	73.06	**0.8293**	**20.88**
VGG and attr. loss	72.18	59.59	73.51	73.72	73.62	0.8239	20.65
VGG and attr. loss (+ input attr.)	**81.1**	**74.8**	**84.29**	**85.61**	**84.94**	0.8274	20.7
Occlusion	65.74	51.06	68.72	64.36	66.47	0.7153	14.57
GT data	78,66	66,23	77.85	79.71	78.77	-	-

Table 7. Ablation study results on AiC dataset

Method	mean Accuracy	Accuracy	Precision	Recall	F1	SSIM	PSNR
Baseline	72.72	45.48	48.23	80.87	60.42	0.6236	20.49
VGG loss	78.12	53.11	55.52	85.65	67.37	0.7088	21.5
VGG and attr. loss	78.37	53.3	55.73	85.46	67.46	**0.7101**	**21.81**
VGG and attr. loss (+ input attr.)	**90.86**	**72.15**	**74.0**	**95.1**	**83.23**	0.6986	21.47
Occlusion	72.24	45.77	48.78	79.03	60.32	0.6148	18.38
GT data	91.89	74.87	76.80	95.43	85.11	-	-

Table 8. Comparison with the state-of-the-art method on RAP dataset

Method	mA	Accuracy	Precision	Recall	F1	SSIM	PSNR
Pix2Pix [21]	69.49	52.05	65.07	70.06	67.47	0.7348	17.91
RN [14]	65.92	51.44	65.77	67.94	66.84	0.6798	18.4
Ours	**72.18**	**59.59**	**73.51**	**73.72**	**73.62**	**0.8239**	**20.65**
Occlusion	65.74	51.06	68.72	64.36	66.47	0.7153	14.57

In order to generate people without occlusion, a classical adversarial loss is employed where a discriminator is tasked to distinguish between real and fake fully visible people. To generate images that have similar feature representations we adopted a perceptual loss [22]. Rather than encouraging the pixels of the output image to exactly match the pixels of the target image, we instead encourage them to have similar feature representations as computed by the VGG16 network. Finally, since our main purpose is not limited to naively restore the occluded parts of pedestrians, but also to maintain and highlight their attributes, we introduced an additional loss component. As for the perceptual loss, we used a classification network as loss function. In particular, we adapted ResNet-101, pre-trained on ImageNet, to the task of multi-attribute classification. Differently, from the VGG loss, we work on a higher level of abstraction, forcing the Generator network to produce images that exhibit characteristics coherent with the attributes of the person. This is another example of a superhuman capability that would never be possible without the help of a synthetic "superhuman" dataset.

Fig. 9 shows some qualitative results. The baseline performs considerably worse than the other experiments, not being able to completely remove the occlusions on AiC. The synthetic dataset is, in fact, more challenging compared to our corrupted version of RAP. For the same reason, RAP results are overall more appealing than the ones obtained on AiC. Moreover, no substantial difference appears between the other setups, highlighting the fact that the VGG loss is the main component that guides the network to produce high-quality results.

Fig. 9. Qualitative results based on the ablation study on RAP dataset (leftmost) and AiC dataset (rightmost). GT columns indicate ground truth images while in the OCC columns are presented the input occluded images. Columns 3 and 9 indicate the outputs of our baseline. Columns 4 and 10 represents results of the VGG loss. On 5 and 11 we have results of experiments using all the 3 losses combined: adversarial loss, VGG loss, and attribute loss. Finally, columns 6 and 12 show results where attributes are injected as input to the network. The Figure is taken from "Can Adversarial Networks Hallucinate Occluded People With a Plausible Aspect?" [17]

Table 6 and Table 7 present quantitative results for RAP and AiC respectively based on our ablation study. The tables also provide metrics referred to the occluded images before the restoration process. Despite being visually indistinguishable, the images obtained from the VGG loss and from our Entire configuration produce very different results in terms of attribute metrics. We can also observe that there is no substantial difference between the VGG loss and the VGG loss with Attributes loss. In fact, RAP shows a gap of one percentage point in almost all the classification metrics, while AiC shows very little differences, due to the more challenging nature of AiC.

Moreover, Table 6 shows that the Entire setup reaches higher scores compared to the upper bound of the ground truth images. Also, Table 7 shows performances that are close to the ground truth metrics when we input attribute information directly to the Generator. In fact, with attributes as input, the Generator network, by restoring the occluded images, is able to produce an output that has enhanced attribute characteristics (although this is not visible to the naked eye).

This is an example of what can be done with Generative Networks tasked to fill the gaps due to occlusions and by creating a fine-grained representation of the human shape. As the matter of fact, this approach could be improved by a deeper exploration of the best architectures to extract human information that is used to produce the supervised signal that guides the training procedure. However, in spite of the choice of the generative architecture (the U-net in our example) and the discriminative networks (VGG-16, ResNet-101, and a two-class CNN), the lesson learned is that the goal of a good design is to match the embedding capabilities with the specific task. The compressed neural representation of the body shape learned by the generative network is conditioned by an estimated knowledge, i.e., the attribute vector of the shape. Indeed, the reconstructive capability of the network can be compared with a human-like imagination, which is equally affected by biases.

Table 9. Comparison on the CMU Panoptic dataset. Results are shown in terms of MPJPE [mm] and F1 detection score. Last row: results with ground truth volumetric heatmaps

	Haggl.	Mafia	Ultim.	Pizza	Mean	F1
Rogez *et al.* [31]	218	187	194	221	203	-
Zanfir *et al.* [53]	140	166	151	156	153	-
Zanfir *et al.* [54]	72	**79**	67	94	72	-
LoCO	**45**	95	**58**	**79**	**69**	89.21
GT	9	12	9	9	10	100

Table 10. Comparison on the Human3.6m dataset in terms of average MPJPE [mm]. "(a)" indicates the addition of rigid alignment to the test protocol; N is the number of joints considered by the method. "TD" and "BU" indicates top-down and bottom-up methods respectively. Last row: results with ground truth volumetric heatmaps

	Method	N	P1	P1 (a)	P2	P2 (a)
TD	Rogez *et al.* [36]	13	63.2	53.4	87.7	71.6
	Debral *et al.* [10]	16	-	-	-	65.2
	Rogez *et al.* [38]	13	54.6	45.8	65.4	54.3
	Moon *et al.* [30]	17	**35.2**	**34.0**	**54.4**	53.3
BU	Mehta *et al.* [28]	17	-	-	80.5	-
	Mehta *et al.* [27]	17	-	-	69.9	-
	LoCO	14	51.1	43.4	61.0	**49.1**
GT Vol. Heatmaps		14	15.6	14.9	15.0	14.3

This network could be used as a support for Multimedia Surveillance, forensics, or security-related applications in order to give more information about the appearance of a person that has been acquired under severe occlusion. Finally, it is important to note that the reconstruction ability of the network is dependent on the fairness of the training dataset as biases on the dataset could distort the results considerably.

5 HUMAN 3D ASSESSMENT BY SUPERVISED POSE LEARNING

A third example of what can be generated artificially by a network is the 3D estimation of the spatial distribution of people in surveillance scenes. Humans are not able to accurately predict the distance of objects and persons by simply looking at them. In few meters, humans can estimate distances by relying on their stereo vision. Exceeded this distance, humans use learned perspective information to infer the 3D distances, in accordance with our long-lasting visual experience. Similarly, surveillance systems are able to do the same thanks to Machine Learning. The goal of estimating three-dimensional human positions and pose by solely relying on monocular images is a very new and challenging task in Computer Vision, that has been recently tackled in a top-down manner by firstly detecting the target people and then estimating the distance of the joints of a single person w.r.t. the camera location.

A more efficient approach exploits bottom-up supervised learning approaches trained on synthetic data which outputs a 3D pose estimation for every person in the image in a single forward pass, as proposed by the Learning on Compressed Output (LoCO) architecture [15].

In LoCO we infer the localization of every person starting from an estimation of the 3D position of all the detected joints in an image. Thus, the basic idea is to predict the 3D positions of all heads, knees, feet, etc., and then to group them into a skeleton by relying on some (learned) physical constraints of the human body. For instance, we learn that

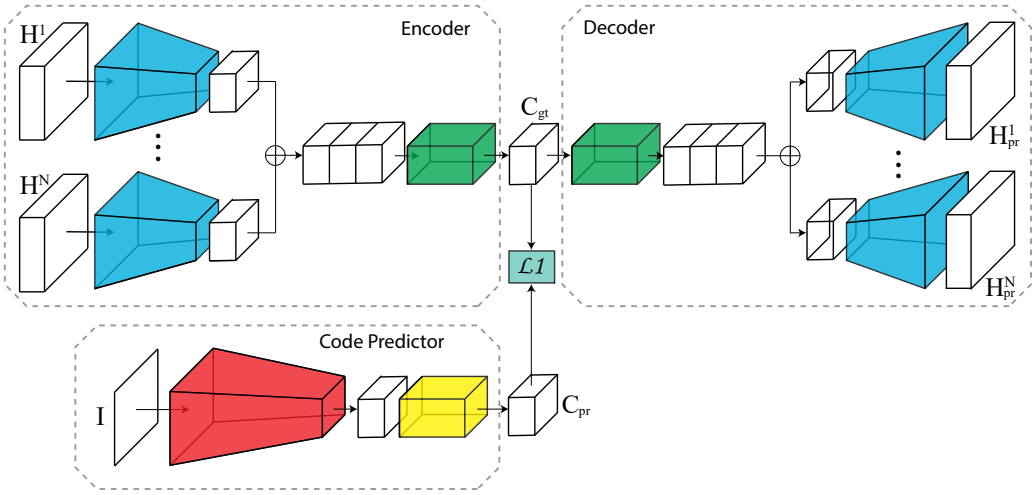

Fig. 10. Schematization of the LoCO pipeline. At training time, the Encoder takes the Volumetric Heatmaps $H^1, ..., H^N$ and produces the compressed volumetric heatmaps C_{gt} which are used as ground truth from the Code Predictor. At test time, the intermediate representation computed by the Code Predictor is fed to the Decoder for the final output.

the distance between hands or between head and feet is limited and changes in accordance with the distance due to the perspective constraints.

Specifically, LoCO is an approach for bottom-up multi-person 3D human pose estimation from monocular RGB images which models joint location with high-resolution volumetric heatmaps, devising a simple and effective compression method to drastically reduce the size of this representation. At the core of the method lies the Volumetric Heatmap Autoencoder, a fully-convolutional network tasked with the compression of ground-truth heatmaps into a dense intermediate representation. Figure 10 shows a schematization of the LoCO pipeline.

A second model, the Code Predictor, is then trained to predict these codes, which can be decompressed at test time to re-obtain the original representation. The experimental evaluation shows that this method performs favorably when compared to state of the art on both multi-person and single-person 3D human pose estimation datasets and, thanks to the novel compression strategy, can process full-HD images at the constant run-time of 8 fps regardless of the number of subjects in the scene.

The core of the proposal relies on the creation of an alternative ground-truth representation that preserves the most informative content of the original ground-truth but reduces its memory footprint. Indeed, this new compressed representation is used as the target ground-truth during our network training. By leveraging on the analogy between compression and dimensionality reduction on sparse signals [3, 41, 46], we empirically follow the intuition that 3D body poses can be represented in an alternative space where data redundancy is exploited towards a compact representation. This is done by minimizing the loss of information while keeping the spatial nature of the representation, a task for which convolutional architectures are particularly suitable. Concurrently w.r.t. LoCO, compression-based approaches have been effectively used for both dataset distillation and input compression [44, 47] but, to the best of our knowledge, this is the first time they are applied to ground truth remapping. For this purpose, deep self-supervised networks such as autoencoders represent a natural choice for searching, in a data-driven way, for an intermediate representation.

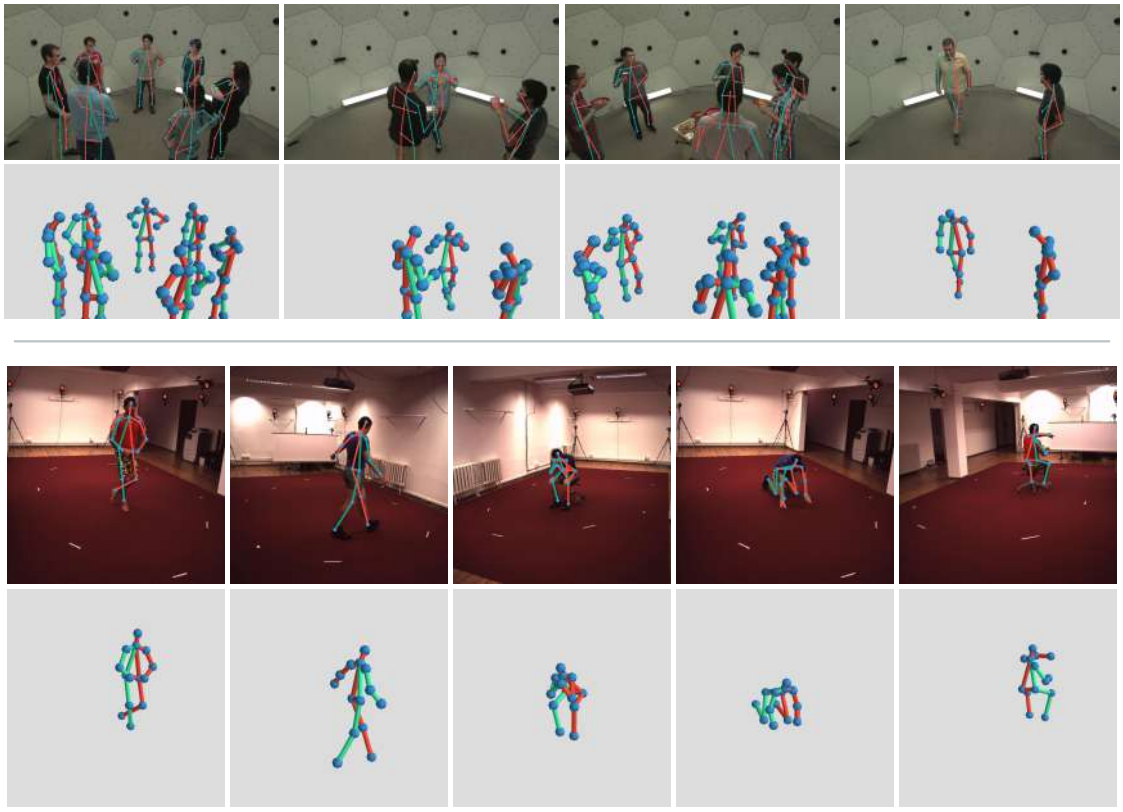

Fig. 11. Qualitative results of our LoCO approach. 1st and 2nd rows: result of LoCO on the CMU Panoptic dataset; 3rd and 4th rows: result of LoCO on the Human3.6m dataset. The Figure is taken from "Compressed Volumetric Heatmaps for Multi-Person 3D Pose Estimation" [15].

Our LoCO approach allows us to exploit Volumetric Heatmaps as a ground truth representation for the 3D pose estimation task. Instead, without compression, this would lead to a sparse and extremely high dimensional output space with consequences on both the network size and the stability of the training procedure. In comparison with top-down approaches, we removed the dependency on the people detector stage, hence gaining both in terms of robustness and assuring a constant processing time at the increasing of people in the scene. The experiments show state-of-the-art performance on all the considered datasets. Figure 11 shows some qualitative results of the method.

Results obtained on the CMP Panoptic dataset are shown in Table 9, divided by action type and expressed in terms of Mean Per Joint Position Error (MPJPE). The obtained results show the advantages of using volumetric heatmaps for 3D Human Pose Estimation, as LoCO achieves the best result. Table 10 further shows a comparison with state-of-the-art multi-person methods on the single person Human3.6M dataset, showing that LoCO is well suited even in the single person context, as it achieves state of the art results among bottom-up methods.

This can be considered a mixed approach taking the pros and cons of the two methods previously discussed. Assuming that 2D pose estimation has acceptable results and that extracting fine-grained information (i.e., the joints) without appearance information prevents having misleading information that is distorted by perspective, we use a network to

Fig. 12. GUI of our system. In the main frame, anonymized bounding boxes are superimposed to the image. Colored links encodes people reciprocal distance. On the right, two maps shows the bird-eye view of the area. The estimated risk level of the scene resides at the bottom of the interface.

hallucinate not the pixel-level appearance but the skeleton. In this manner, every joint can be reconstructed at the same time in the 3D space, with very high confidence even when overlapped and occluded.

6 AN EXAMPLE OF APPLICATION

The previously discussed datasets can be also exploited for real-world applications. In fact, we utilized our synthetic dataset to benchmark an in-edge AI system designed to monitor the acceptance of social distancing prevention measures during the COVID-19 pandemic. The proposed system can model the risk of possible contagious in a given area monitored by RGB cameras where people freely move and interact. The system, called Inter-Homines, evaluates in real-time the contagion risk by analyzing video streams: it is able to locate people in 3D space, calculate interpersonal distances and predict risk levels by building dynamic maps of the monitored area. The system has been tested on our synthetically generated datasets. Despite being synthetic, our data features highly challenging and complex situations, peculiar of surveillance scenarios, where people are often dominated by severe body part occlusions and truncation. For those reasons, we believe this data is the perfect choice to validate a system that targets global safety.

The system has a twofold goal. The first is to provide a reliable tool, in accordance with European privacy and usage guidelines of the AI, to calculate in real-time the actual compliance with the prevention measures for "people spacing", also interactively reporting any risky situations. In particular, the implemented system can generate real-time alarms when people form crowds. The second goal is to provide an innovative model for the dynamic calculation of the risk of the monitored site that can be used as a tool for prevention, control, monitoring, and planning, support to the population and workers in order to implement conscious attendance, linked to effective compliance with the measures in force. The aim of our Inter-Homines system is to detect people, compute their distance and provide a dynamic risk

Fig. 13. Examples of CenterNet bounding boxes (pink), refined bounding boxes and head localization (green).

level of the area, as well as producing a human-readable visualization with anonymized people. For GDPR constraints, no visual data is recorded but, instead, only people coordinates are extracted and stored. Data is acquired with a variable rate, up to one time per second for each camera. Figure 12 shows the graphic user interface of the application.

In this project, we provide a novel detection pipeline running in real-time. It exploits standard fast camera calibrations, a people detector, and pose estimation methods. As we are interested in the best speed-accuracy trade-off, we choose CenterNet [56] as a people detector which yields 51.3% AP for the people class on MS COCO, running at 52 FPS on a Titian XP. CenterNet is capable of producing a precise localization of every person in the image, however, it does not take into account occlusions that usually happen in real-world scenarios. If a person is occluded by an object or by other people, CenterNet predicts a tight bounding box that only contains the visible part of the person, ignoring his full shape. This usually happens with the bottom part of the body, as the camera is commonly placed several meters above the ground. Since we are ultimately interested in recovering the ground plane coordinate of each person through homography, we need to know the exact position (in the image plane) of the feet of each detected person. This task cannot be accomplished by solely relying on CenterNet.

To overcome the aforementioned limitations without introducing complexity to the overall system, we propose to utilize a small network to predict the feet position given a bounding box containing a person, even if the feet are not visible. To this aim we rely on a simple modification of THOPA-Net, given an image tightly containing a person, to regress to the midpoint of the segment having the two feet as endpoints. This ensures that we know the exact position in the image plane where every person touches the ground. Figure 13 shows some examples of refined bounding boxes. Since we are also interested in anonymizing the face of each detected person, we further predict the location of the head, by the same network.

For this module, we used JTA as the training dataset since it is the only surveillance dataset available in the literature that provide pose estimation annotations with occlusion information. Thanks to this, we were able to simulate occlusion situations by simply picking, during training, the pedestrians with the bottom keypoints occluded, like ankles, knees, and hips. During training, we also randomly shortened some of the bounding boxes in order to simulate CenterNet behaviors. This step ensures a more precise localization of the feet while also coping with truncated bounding boxes. Our network can effectively obtain an accurate position of each head and it is used to extend the bounding box to its

regular shape. In this application, we do not exploited the LoCO estimation of 3D joints since we can rely on fully calibrated cameras to infer the distances between people.

7 CONCLUSIONS AND ACKNOWLEDGE

This paper discussed some ideas for 2D and 3D people detection for surveillance applications, with a specific focus on occlusion. Having Artificial Intelligence modules capable of estimating the human pose in the space also under severe occlusions and perspective size deformation allows surveillance systems to reach some superhuman vision capabilities, that can be exploited in multimedia interfaces to empower human capabilities in monitoring and control. This can be exploited both in real-time to assess dangerous situations or for forecasting and statistical evaluation of the context, as in the case of risk assessment for contagiousness in monitored areas. Nowadays, neural architectures are becoming effective in supervised and semi-supervised related tasks thanks to the availability of open-source datasets. These datasets should be rich, collected with fairness, and in accordance with the values of equity (e.g. gender equity) and explainable capabilities. For this reason, the use of simulated environments could be a good answer to such constraints. We would like to acknowledge the researchers at AimageLab who supported and co-authored some of the cited works. The projects discussed in this paper are supported by the Italian Ministry of University and Research under PRIN Project PREVUE (PRediction of Events in Urban Environments) and the European projects ARTEMIS Arrowhead Tools, as well as the NVIDIA AI Technology Center at UNIMORE.

REFERENCES

[1] [n.d.]. White Paper on Artificial Intelligence: Public consultation towards a European approach for excellence and trust. https://ec.europa.eu/digital-single-market/en/news/white-paper-artificial-intelligence-public-consultation-towards-european-approach-excellence. Accessed: 04-02-2021.

[2] Mykhaylo Andriluka, Umar Iqbal, Anton Milan, Eldar Insafutdinov, Leonid Pishchulin, Juergen Gall, and Bernt Schiele. 2018. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5167–5176.

[3] H. Arai, Y. Chayama, H. Iyatomi, and K. Oishi. 2018. Significant Dimension Reduction of 3D Brain MRI using 3D Convolutional Autoencoders. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 5162–5165. https://doi.org/10.1109/EMBC.2018.8513469

[4] S. H. Bae and K. J. Yoon. 2018. 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. https://doi.org/10.1109/TPAMI.2017.2691769

[5] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*.

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7291–7299.

[7] Q. Chu, W. Ouyang, H. Li, X. Wang, B. Liu, and N. Yu. 2017. Online Multi-object Tracking Using CNN-Based Single Object Tracker with Spatial-Temporal Attention Mechanism. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 4846–4855. https://doi.org/10.1109/ICCV.2017.518

[8] Rita Cucchiara. 2005. Multimedia surveillance systems. In *Proceedings of the third ACM international workshop on Video surveillance & sensor networks*. 3–10.

[9] Rita Cucchiara, Costantino Grana, Massimo Piccardi, and Andrea Prati. 2003. Detecting moving objects, ghosts, and shadows in video streams. *IEEE transactions on pattern analysis and machine intelligence* 25, 10 (2003), 1337–1342.

[10] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. 2018. Learning 3D Human Pose from Structure and Motion. In *European Conference on Computer Vision (ECCV)*.

[11] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. 2020. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003* (2020).

[12] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2014. Pedestrian Attribute Recognition At Far Distance. In *Proceedings of the 22Nd ACM International Conference on Multimedia*.

[13] Matteo Fabbri, Guillem Brasó, Gianluca Maugeri, Aljoša Ošep, Riccardo Gasparini, Orcun Cetintas, Simone Calderara, Laura Leal-Taixé, and Rita Cucchiara. 2021. MOTSynth: How Can Synthetic Data Help Pedestrian Detection and Tracking?. In *International Conference on Computer Vision (ICCV)*.

[14] Matteo Fabbri, Simone Calderara, and Rita Cucchiara. 2017. Generative adversarial models for people attribute recognition in surveillance. In *2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS)*. IEEE, 1–6.

[15] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Stefano Alletto, and Rita Cucchiara. 2020. Compressed volumetric heatmaps for multi-person 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7204–7213.

[16] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Andrea Palazzi, Roberto Vezzani, and Rita Cucchiara. 2018. Learning to detect and track visible and occluded body joints in a virtual world. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 430–446.

[17] Federico Fulgeri, Matteo Fabbri, Stefano Alletto, Simone Calderara, and Rita Cucchiara. 2019. Can adversarial networks hallucinate occluded people with a plausible aspect? *Computer Vision and Image Understanding* 182 (2019), 71–80.

[18] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *CVPR*.

[19] Ismail Haritaoglu, David Harwood, and Larry S. Davis. 2000. W/sup 4: real-time surveillance of people and their activities. *IEEE Transactions on pattern analysis and machine intelligence* 22, 8 (2000), 809–830.

[20] Umar Iqbal, Anton Milan, and Juergen Gall. 2017. Posetrack: Joint multi-person pose estimation and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 1.

[21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. 5967–5976. https://doi.org/10.1109/CVPR.2017.632

[22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*. 694–711.

[23] Philipp Krähenbühl. 2018. Free supervision from video games. In *CVPR*.

[24] Dangwei Li, Zhang Zhang, Xiaotang Chen, Haibin Ling, and Kaiqi Huang. 2016. A richly annotated dataset for pedestrian attribute recognition. *preprint arXiv:1603.07054* (2016).

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

[26] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Junjie Yan, and Xiaogang Wang. 2017. HydraPlus-Net: Attentive Deep Features for Pedestrian Analysis. In *Proceedings of the IEEE international conference on computer vision*. 1–9.

[27] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. 2018. Single-shot multi-person 3d pose estimation from monocular rgb. In *International Conference on 3D Vision (3DV)*.

[28] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. 2017. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)* (2017).

[29] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. 2016. MOT16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831* (2016).

[30] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. 2019. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10133–10142.

[31] Alin-Ionut Popa, Mihai Zanfir, and Cristian Sminchisescu. 2017. Deep multitask architecture for integrated 2d and 3d human sensing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[32] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497* (2015).

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*.

[35] Stephan R Richter, Zeeshan Hayder, and Vladlen Koltun. 2017. Playing for benchmarks. In *Proceedings of the IEEE International Conference on Computer Vision*. 2213–2222.

[36] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. 2017. Lcr-net: Localization-classification-regression for human pose. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[37] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. 2019. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *IEEE transactions on pattern analysis and machine intelligence* 42, 5 (2019), 1146–1161.

[38] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. 2019. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

[39] A. Sadeghian, A. Alahi, and S. Savarese. 2017. Tracking the Untrackable: Learning to Track Multiple Cues with Long-Term Dependencies. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 300–311. https://doi.org/10.1109/ICCV.2017.41

[40] Ricardo Sanchez-Matilla, Fabio Poiesi, and Andrea Cavallaro. 2016. Online Multi-target Tracking with Strong and Weak Detections. In *Computer Vision – ECCV 2016 Workshops*, Gang Hua and Hervé Jégou (Eds.). Springer International Publishing, Cham, 84–99.

[41] Matthias Scholz, Martin Fraunholz, and Joachim Selbig. 2008. Nonlinear principal component analysis: neural network models and applications. In *Principal Manifolds for Data Visualization and Dimension Reduction*.

[42] F. Solera, S. Calderara, and R. Cucchiara. 2015. Towards the evaluation of reproducible robustness in tracking-by-detection. In *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 1–6. https://doi.org/10.1109/AVSS.2015.7301755

[43] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. 2020. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*.

[44] Róbert Torfason, Fabian Mentzer, Eiríkur Ágústsson, Michael Tschannen, Radu Timofte, and Luc Van Gool. 2018. Towards Image Understanding from Deep Compression Without Decoding. In *International Conference on Learning Representations*. https://openreview.net/forum?id=HkXWCMbRW

[45] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, B.B.G Sekar, Andreas Geiger, and Bastian Leibe. 2019. MOTS: Multi-Object Tracking And Segmentation. In *CVPR*.

[46] Jing Wang, Haibo He, and Danil V Prokhorov. 2012. A folded neural network autoencoder for dimensionality reduction. *Procedia Computer Science* (2012).

[47] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. 2018. Dataset distillation. *arXiv preprint arXiv:1811.10959* (2018).

[48] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. 2017. Simple Online and Realtime Tracking with a Deep Association Metric. In *2017 IEEE International Conference on Image Processing (ICIP)*. 3645–3649.

[49] Christopher Richard Wren, Ali Azarbayejani, Trevor Darrell, and Alex Paul Pentland. 1997. Pfinder: Real-time tracking of the human body. *IEEE Transactions on pattern analysis and machine intelligence* 19, 7 (1997), 780–785.

[50] Bin Xiao, Haiping Wu, and Yichen Wei. 2018. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*. 466–481.

[51] Fengwei Yu, Wenbo Li, Quanquan Li, Yu Liu, Xiaohua Shi, and Junjie Yan. 2016. POI: Multiple Object Tracking with High Performance Detection and Appearance Feature. In *Computer Vision – ECCV 2016 Workshops*, Gang Hua and Hervé Jégou (Eds.). Springer International Publishing, Cham, 36–42.

[52] Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. 2018. PCN: Point Completion Network.

[53] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. 2018. Monocular 3D Pose and Shape Estimation of Multiple People in Natural Scenes - The Importance of Multiple Scene Constraints. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[54] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu. 2018. Deep network for the integrated 3d sensing of multiple people in natural images. In *Advances in Neural Information Processing Systems*.

[55] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*. 1116–1124.

[56] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. 2019. Objects as points. *arXiv preprint arXiv:1904.07850* (2019).