University of Modena and Reggio Emilia
XXXIII Cycle of the International Doctorate School in
Information and Communication Technologies

Doctor of Philosophy Dissertation in
Computer Engineering and Science

Exploiting Synthetic Data to Improve Human Behavior Understanding

Matteo Fabbri

Supervisor: Prof. Rita Cucchiara
PhD Course Coordinator: Prof. Sonia Bergamaschi

Modena, 2021

Review committee composed of:

Dr. Fabrizio Falchi
Institute for Information Science and Technologies
"Alessandro Faedo" - CNR

Dr. Stefano Alletto
Apple Inc.

To Fabio

A bad design with a good presentation is doomed eventually. A good design with a bad presentation is doomed immediately.

Akin's Laws of Spacecraft Design

Abstract

Most recent Deep Learning techniques require large volumes of training data in order to achieve human-like performance. Especially in Computer Vision, datasets are expensive to create because they usually require a considerable manual effort that can not be automated. Indeed, manual annotation is error-prone, inconsistent for subjective tasks (e.g. age classification), and not applicable to particular data (e.g. high frame-rate videos). For some tasks, like pose estimation and tracking, an alternative to manual annotation implies the use of wearable sensors. However, this approach is not feasible under some circumstances (e.g. in crowded scenarios) since the need to wear sensors limits its application to controlled environments. To overcome all the aforementioned limitations, we collected a set of synthetic datasets exploiting a photorealistic videogame. By relying on a virtual simulator, the annotations are error-free and always consistent as there is no manual annotation involved. Moreover, our data is suitable for in-the-wild applications as it contains multiple scenarios and a high variety of people appearances. In addition, our datasets are privacy compliant as no real human was involved in the data acquisition. Leveraging this newly collected data, extensive studies have been conducted on a plethora of tasks. In particular, for 2D pose estimation and tracking, 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. For 3D pose estimation, we propose to use high-resolution volumetric heatmaps to model joint locations, devising a simple and effective compression method to drastically reduce the size of this representation. For attribute classification, we overcome a common problem in surveillance, namely people occlusion, by designing a network capable of hallucinating occluded people with a plausible aspect. From a more practical point of view, we design an edge-AI system capable of evaluating in real-time the COVID-19 contagion risk of a monitored area by analyzing video streams. As synthetic data might suffer domain-shift related problems, we further investigate image translation techniques for the tasks of head pose estimation, attribute recognition and face landmark localization.

Abstract (Italian)

Le più recenti tecniche di Deep Learning richiedono enormi quantità di dati di addestramento per ottenere prestazioni simili a quelle umane. Soprattutto in Computer Vision, i Dataset sono costosi da creare in quanto richiedono uno sforzo manuale considerevole che non può essere automatizzato. Infatti, l'annotazione manuale è spesso soggetta ad errori, è incoerente per task soggettivi (ad es. age classification) e non è applicabile ad ogni tipo di dato (ad es. video ad elevato frame rate). Per alcuni task, come la pose estimation e il tracking, un'alternativa all'annotazione manuale implica l'utilizzo di sensori indossabili. Tuttavia, questo approccio non è praticabile in alcune circostanze (ad es. in scenari affollati), poiché la necessità di indossare tali sensori limita la sua applicazione ad ambienti controllati. Per superare questi limiti, abbiamo raccolto una serie di dati sintetici sfruttando un videogioco fotorealistico. Grazie all'utilizzo di un simulatore virtuale, le annotazioni sono prive di errori e sempre coerenti dato che non sono coinvolte operazioni manuali. Inoltre, i nostri dati sono adatti per applicazioni in-the-wild in quanto contengono un'elevata varietà di scenari e persone in ambienti non controllati. Tali dati sono conformi alle normative sulla privacy, in quanto nessun essere umano è stato coinvolto nell'acquisizione dei video. Sfruttando questi nuovi dati, sono stati condotti studi approfonditi su una serie di task. In particolare, per la pose estimation 2D e il tracking, abbiamo sviluppato un'architettura Deep che estrae congiuntamente i giunti delle persone e le associa su brevi intervalli temporali. Il nostro modello è in grado di ragionare esplicitamente riguardo a parti del corpo occluse, proponendo soluzioni plausibili di giunti non visibili. Per la pose estimation 3D, invece, abbiamo scelto di utilizzare heatmap volumetriche ad alta risoluzione per modellare le posizioni dei giunti, ideando un metodo di compressione semplice ed efficace per ridurre drasticamente le dimensioni di questa rappresentazione. Per l'attribute classification, abbiamo proposto una soluzione ad un problema comune nell'ambito della videosorveglianza, ovvero l'occlusione delle persone, progettando una rete neurale in grado di generare porzioni di persone occluse con un aspetto plausibile. Da un punto di vista pratico, abbiamo progettato un sistema di edge-AI in grado di valutare in tempo reale il rischio di contagio COVID-19 di un'area monitorata analizzando flussi video. Poiché i dati sintetici potrebbero essere suscettibili al domain-shift, abbiamo approfondito le tecniche di image-translation per head pose estimation, attribute recognition e face landmark localization.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Deep learning-based methods require large volumes of training data to achieve good performance. However, data acquisition and annotation for computer vision applications usually demand a substantial amount of manual effort, especially in the video domain. This poses a significant problem, as data acquisition in crowded public environments raises data privacy concerns as we are not allowed to simply record and store data without explicit consent of all participants. Furthermore, labeling instances of pedestrians in highly crowded-scenarios is very challenging even for human annotators and may introduce errors in the training data. In this thesis, we take the performance of pedestrian detection, multi-object tracking, pose estimation and attributes recognition methods to the next level by generating large, highly diverse synthetic datasets using a photo-realistic rendering game-engine. This way, we can simulate highly-crowded and diverse environments with perfect annotations.

The newly generated data enables a set of new challenging tasks that are not feasible by solely relying on manually annotated datasets. Specifically, in Section 3.1, 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. The 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. Indeed, temporal continuity in the detection phase gains more

importance when scene cluttering introduces the challenging problems of occluded targets.

In Section 3.2 we present a novel approach for bottom-up multi-person 3D human pose estimation from monocular RGB images. We propose to use high resolution volumetric heatmaps to model joint locations, devising a simple and effective compression method to drastically reduce the size of this representation. At the core of the proposed method lies the Volumetric Heatmap Autoencoder, a fully-convolutional network tasked with the compression of ground-truth heatmaps into a dense intermediate representation. 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.

From a more practical perspective, in Section 3.3 we utilize our synthetic datasets 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 contagiousity 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 truncations. For those reasons, we believe this data is the perfect choice to validate a system that targets global safety.

Finally, in Section 3.4, we design a network capable of generating a complete image of a person, given an occluded version in input. The generated image should depict a fully visible person similar to a completely visible people shape and able to conserve similar visual attributes of the original one. For the purpose, we propose a new approach by integrating the state-of-the-art of Generative Adversarial Networks (GAN) as well as discriminative attribute classification nets, with an architecture specifically designed to de-occlude people shapes. This work could be an initial step to many further researches to recognize people and their behavior in an open crowded world.

Deep learning methods trained on synthetically generated data usually suffer domain-shift related problems. For this reason, we investigate domain adaptation techniques in order to bridge the gap between "source domain" and "target domain" for head pose estimation, attributes recognition and facial landmark localization. In particular, in Section 4.1, we propose a complete framework for the estimation of the head and shoulder pose relying on depth images only where a Face-from-Depth component based on a Conditional GAN is able to hallucinate a face from the corresponding depth image. In addition to a performance improvement, the introduction of the Face-from-Depth module allows us to train the system on wider datasets since more annotated data on gray-level images are usually available rather than on depth ones.

Additionally, in Section 4.2, we further explore the capabilities of the Face-from-Depth component. Although the network cannot reconstruct the exact somatic features for unknown individual faces, it is capable of reconstructing plausible faces as their appearance is accurate enough as it can be used in multiple pattern recognition tasks. In fact, we test the network capability to hallucinate plausible faces with two perceptual probes: face attributes classification and landmark localization. Experimentally, we demonstrate that this domain translation technique can constitute a new way of exploiting depth data in advanced future applications.

Chapter 2

Literature Survey

In the following sections we briefly report other research approaches that have tackled topics related to this thesis. The list of methods could be much longer but we chose to restrict ourselves to the ones most relevant for the community and the ones most relevant to us because of similarities with the proposed algorithms.

2.1 Datasets

Advances in computer vision have been driven by the constant growth of available datasets and benchmarks. The Pascal VOC [59] was instrumental in the progress of deep neural networks for object detection. The ImageNet [230] dataset supported the development of visual classification technique that broadly influenced the field. Microsoft COCO [149] support research on semantic instance segmentation and object detection while the SUN [275] and Places [302] datasets provides data for scene recognition.

Multi-Object Tracking.

In the tracking community, one of the pioneers is the KITTI benchmark [74], which maintains ground truth for object detection and tracking. Sequences are collected by a camera mounted on a car moving through traffic. However, bounding box annotation is available for a small number of frames and data only represents the appearance of a single city in clear weather. nuScenes [26] is a newly

released large-scale driving dataset that also provides object detection and tracking annotation of people. It features sequences recorded in two different cities with varying illuminations and weather conditions. However, like KITTI, nuScenes is an autonomous driving dataset and lacks crowded scenarios specific of surveillance contexts. The H3D benchmark [203] tries to cope with low pedestrian density by providing ground truth for highly interactive and occluded traffic participants. Yet, the challenges introduced are not comparable to the ones peculiar of the most crowded areas, like airports, stations, mall or city centers.

Pedestrian Tracking

In the last few years, the MOTChallenge [47] suite played a pivotal role in the improvement of the most recent multi-object tracking techniques by introducing clean datasets and a precise framework. In particular, MOT17 [176] is the reference benchmark for the evaluation of multi-person tracking for surveillance purposes. It provides challenging sequences of crowded urban scenes with severe occlusions and scale variations. MOT20 [48] is the latest benchmark of the MOTChallenge suite which has been specifically designed to push the limits of the emerging techniques when it comes to handling extremely crowded scenes.

People tracking is deeply entwined with person re-identification, as the most successful tracking approaches showed the importance of learned reID features [22]. Among the publicly available datasets that provide ID annotation, Market1501 [300], CUHK03 [143] and DukeMTMC [223] are the most commonly used by the tracking community.

Synthetic Data

Collecting data usually demand a tremendous amount of work as it involves a series of manual procedures that can not be easily automated. Indeed, creating a dataset often requires raw data collection, annotation and error check strategies in a strict and well-defined protocol. As more data is constantly required to train ever growing models, the cost of such datasets is becoming prohibitive. This burden can either limit the quality or the quantity of data acquired, slowing down progress in Computer Vision.

A possible solution to the aforementioned problems is to employ virtual worlds. Simulated environments have been used to evaluate optical flow algorithms [10, 25, 221, 169, 127], depth estimation [127] visual odometry systems [89, 91, 298, 221] and to benchmark the robustness of feature descriptors [118]. Simulated worlds

have been utilized to test visual surveillance systems [254], evaluate multi-object tracking [73, 63, 127], hand tracking models [240], human pose estimation [242, 63, 82] and crowd counting [266]. Virtual environments have also been applied to pedestrian detection [163, 2], stereo reconstruction [170], and semantic segmentation [90, 229, 100, 221, 127, 222].

However, none of the previous attempts of creating synthetic datasets were able to completely replace real data. In fact, the majority of simulated datasets are only used to benchmark and validate new techniques, but they fail when models trained on that data are applied to real world contexts. In fact, domain adaptation techniques [21] are still required to effectively bridge the gap between synthetic and real worlds. In this thesis, we go beyond this, by showing that synthetic data, when varied enough, can be used as a full proxy for real world applications, without having to rely on fine-tuning or domain adaptation techniques.

2.2 Human Behaviour Understanding

In this section we briefly report other research approaches closely related to Human Behaviour Understanding for the tasks of people detection, 2D pose estimation and tracking, 3D pose estimation, head detection, head pose estimation and attribute recognition.

People Detection

One of the most popular two-stage deep object detectors is R-CNN [78] which predicts object location from a set of region proposals [273], crops them and classifies each using a second deep neural network. Fast R-CNN [77], instead, directly crops image features to save computation. However, both approaches rely on slow low-level region proposal techniques.

On the other hand, one-stage methods such as Faster R-CNN [217] generates region candidates within the detection network. It samples bounding boxes with fixed shape (anchors) around the image grid and classifies them into foreground or background. Each proposal is then further classified into object classes. Several improvements to one-stage detectors include anchor shape priors such as in YOLO [211, 212], SSD's different feature resolution [151], and loss re-weighting among different samples [148].

Our people detection approach leverages CenterNet [304], which is closely related to anchor based one-stage detectors. However, CenterNet does not require

manual thresholds for foreground and background classification and does not require Non-Maximum Suppression (NMS) [17] post processing as it simply extracts local peaks in the keypoint heatmap [29, 63]. Moreover, CenterNet utilizes an output stride of 4 which is 2 times larger than in traditional object detectors [94, 93], making it more accurate.

Other methods utilize the same robust keypoint estimation network as CenterNet: CornerNet [134] and ExtremeNet [305]. CornerNet detects the bounding box corners as keypoints while ExtremeNet predicts the left, top, right and bottom extremes of the objects. However, those methods require a combinatorial grouping stage as post processing, which considerably slows down the whole pipeline. CenterNet, instead, simply extracts a single center point per object without the need for grouping or post-processing.

People detection can be also achieved by pose estimation. The trend of pose estimation [29, 186, 103] is very promising, but often is too computationally severe to be implemented for real time edge applications with an unknown number of people. Thus, in more practical applications, we adopt a simplified pose estimation algorithm, that it is used together with the people detector to make the localization more robust to occlusions.

Many 3D object detection methods have been proposed in literature. Among them, 3D R-CNN [129] adds a further head to Faster R-CNN [217] which is followed by a 3D projection. Also Deep Manta [31] exploits a coarse-to-fine Faster R-CNN [217] trained on multiple tasks. Finally, Deep3Dbox [180] utilizes a slow R-CNN [78] by first predicting 2D bounding boxes and then feeding each detection into a 3D estimation network. However, those methods require huge computational power and does not leverage constraints such as fixed camera and flat ground plane.

2D Pose Estimation and Tracking

2D Pose Estimation. Human pose estimation in images has made important progress over the last few years [30, 270, 99, 186, 24]. 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 [208, 103, 105, 198, 185, 29, 138]. Among them, [198] uses graph decomposition and node labeling with local search while [185] 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 [29] while [198], instead, exploits a two-stage approach, consisting of a person detection stage followed by a keypoint estimation for each person. Moreover, [208, 103, 105] 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, [110, 293, 207, 81]. 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.

Tracking. In recent years, online tracking has been successfully extended to scenarios with multiple targets [287, 272, 231, 40, 8, 234]. 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. [287] exploits a high-performance detector with a deep learning appearance feature while [231] presents an online method that encodes long-term temporal dependencies across multiple cues. [40], on the other hand, introduces spatial-temporal attention mechanism to handle the drift caused by occlusion and interaction among targets. [8] solves the online multi-object tracking problem by associating tracklets and detections in different ways according to their confidence values and [234] exploits both high and low confidence target detections in a probability hypothesis density particle filter framework.

Joint Learning. In this thesis, 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 [5, 109] do not tackle pose estimation and tracking simultaneously, but rather target on multi-person tracking alone. More recent methods [106, 102], which rely on graph partitioning approaches closely related to [208, 103, 105], 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 [106, 102] 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.

3D Pose Estimation

Single-Person 3D HPE. Single person 3D HPE from a monocular camera has become extremely popular in the last few years. Literature can be classified into three different categories: (i) approaches that first estimate 2D joints and then project them to 3D space, (ii) works that jointly estimate 2D and 3D poses, (iii) methods that learn the 3D pose directly from the RGB image.

The majority of works on single person 3D HPE first compute 2D poses and leverages them to estimate 3D poses, either using off-the-shelf 2D HPE methods [136, 98, 166, 171, 18, 179, 34] or by having a dedicated module in the 3D HPE pipeline [189, 204, 147, 284].

Joint learning of 2D and 3D pose is also shown to be beneficial [173, 43, 283, 303, 255, 191, 117, 206], often in conjunction with large-scale datasets that only provide 2D pose ground-truth and exploiting anatomical or structure priors.

Finally, recent works estimate 3D pose information directly [236, 205, 250, 158, 188, 219, 220]. Among these, Pavlakos *et al.* [205] were the first to propose a fine discretization of the 3D space around the target by learning a coarse-to-fine prediction scheme in an end to end fashion.

Multi-Person 3D HPE. To the best of our knowledge, very few works tackle multi-person 3D HPE from monocular images. We can categorize them into two classes: top-down and bottom-up approaches. Top-down methods first identify bounding boxes likely to contain a person using third party detectors and then perform single-person HPE for each person detected. Among them, Rogez *et al.* [226] classifies bounding boxes into a set of K-poses. These poses are scored by a classifier and refined using a regressor. The method implicitly reasons using bounding boxes and produces multiple proposals per subject that need to be accumulated and fused. Zanfir *et al.* [290] combine a single person model that incorporates feed-forward initialization and semantic feedback, with additional constraints such as ground plane estimation, mutual volume exclusion, and joint inference. Dabral *et al.* [43], instead, propose a two-staged approach that first estimates the 2D keypoints in every Region of Interest and then lifts the estimated keypoints to 3D. Finally, Moon *et al.* [178] predict absolute 3D human root localization, and root-relative 3D single-person for each person independently. However, these methods heavily rely on the accuracy of the people detector and do not scale well when facing scenes with dozens of people.

In contrast to top-down approaches, bottom-up methods produce multi-person joint locations in a single shot, from which the 3D pose can be inferred even under

strong occlusions. Mehta *et al.* [172], predict 2D and 3D poses for all subjects in a single forward pass regardless of the number of people in the scene. They exploit occlusion-robust pose-maps that store 3D coordinates at each joint 2D pixel location. However, their 3D pose read-out strategy strongly depends on the 2D pose output which makes it limited by the accuracy of the 2D module. Their method also struggles to resolve scenes with multiple overlapping people, due to the missing 3D reasoning in their joint-to-person association process. Zanfir *et al.* [291], on the other hand, utilize a multi-task deep neural network where the person grouping problem is formulated as an integer program based on learned body part scores parameterized by both 2D and 3D information. Similarly to the latter, our method directly learns a mapping from image features to 3D joint locations, with no need of explicit bounding box detections or 2D proxy poses, while simultaneously being robust to heavy occlusions and multiple overlapping people.

Multi-Person 3D Pose Representation. In a top-down framework, the simplest 3D pose representation can be expressed by a vector of joints. By casting 3D HPE as a coordinate regression task, Rogez *et al.* [226] and Zanfir *et al.* [290] indeed utilize x, y, z coordinates of the human joints w.r.t. a known root location. On the other hand, bottom-up approaches require a representation whose coding does not depend on the number of people (e.g. an image map). Among the most recent methods, Mehta *et al.* [172] and Zanfir *et al.* [291] both utilize a pose representation composed by joint-specific feature channels storing the 3D coordinate x, y, or z at the joint/limb 2D pixel location. This representation, however, suffers when multiple overlapping people are present in the scene. In contrast to all these approaches, we adopted the volumetric heatmap representation proposed by Pavlakos *et al.* [205], overcoming all the limitations that arise when facing a multi-person context.

Head Detection.

With RGB or intensity images Viola and Jones [262] face detector is often exploited, *e.g.* in [75, 27, 216, 11, 239]. A different approach demands the head location to a classifier, *e.g.,* [260]. As reported in [174], these approaches suffer due to the lack of generalization capabilities of exploited models, with different acquisition devices and scene contexts.

Recently, deep learning approaches trained on huge face datasets allowed to reach impressive results [282, 29]. However, very few works in literature propose

methods for head detection or localization using *only* depth images as input. A method based on a novel head descriptor and an LDA classifier is described in [37]. Every single pixel is classified as head or non-head, and all pixels are clustered for final head detection. In [187] a fall detection system is proposed, in which is included a module for head detection. Heads are detected only on moving objects through a background suppression. In [65] patches extracted from depth images are used to both compute the location and the pose of the head, through a regression forest algorithm.

Head Pose Estimation

Head pose estimation approaches can rely on different input types: RGB images, depth maps, or both. For this reason, in order to discuss related works, we adopt a classification based on the input data types leveraged by each method.

RGB. RGB methods take monocular or stereo intensity images as input. In [271] a discriminative approach to frame-by-frame tracking the head pose is presented, based on the detection of the centers of both eyes, the tip of the nose and the center of the mouth. Also, [261, 281, 168] leverage well visible facial features on RGB input images, and [251] on 3D data. [56] proposed to predict pose parameters from high-dimensional feature vectors, embedding a Gaussian mixture of linear inverse-regression model into a dynamic Bayesian model. However, these methods need facial (*e.g.* nose and eyes) or pose-dependent features, that should be always visible: consequently, these methods fail when such features are not detected.

A different approach for head pose estimation involves 3D model registration techniques. Firstly, Blanz and Vetter [15] propose a technique for modeling textured 3D faces automatically generated from one or more photographs. Cao *et al.* [28] exploited a 3D regression algorithm that learns an accurate, user-specific face alignment model from an easily acquired set of training data, generated from images of the user performing a sequence of predefined facial poses and expressions. Furthermore, [248] proposed a hybrid approach, which exploits the flexibility of a generative 3D facial model in a combination with a fitting algorithm. However, those techniques often need a manual initialization which is indeed critical for the effectiveness of the method.
A first attempt to use deep learning techniques combined with the regression task in head the pose estimation problem has been performed by Ahn *et al.* [1], through a CNN trained on RGB input images. Also, [193] exploits a CNN by

mapping images of faces on a low dimensional manifold parameterized by pose. In [276] a framework to jointly estimate the head pose and the face alignment using global and local CNN features has been presented while a hybrid approach based on CNN and Gaussian mixture was proposed in [133] and [122]. With deep learning-based approaches, synthetic datasets were often used to train CNNs, that generally require a huge amount of data [152].

Additionally, a bunch of methods regard head pose estimation as an *optimization* problem: in [12] a multi-template, *Iterative Closest Point* (ICP) [156] based gaze tracking system is introduced. Besides, other works use linear or nonlinear regression with extremely low-resolution images [35]. HOG features and a Gaussian locally-linear mapping model were used in [55] and, finally, recent works produce head pose estimations performing a face alignment task [308] using CNNs.
In general, RGB based methods are highly sensitive to illumination, partial occlusions and bad image quality [182].

Depth. Those methods, on the other hand, exploit only range data to perform the pose estimation task. A first attempt to localize accurate nose locations from depth maps in order to perform head tracking and pose estimation was done in [161]. Consequently, [23] used geometric features to identify nose candidates to produce the final pose estimation. A more robust approach was done in [65, 66, 64], where a Random Regression Forest [146] algorithm is exploited for both head detection and pose estimation purposes. Furthermore, in [199] facial point clouds were matched with pose candidates, through a novel triangular surface patch descriptor. As previously stated for RGB methods, those techniques require facial attributes, thus are prone to errors when such features are not detected.

Remaining depth methods regard the head pose estimation task as an *optimization* problem: [195] used the *Particle Swarm Optimization* (PSO) [120] while [174] perform pose estimation by registering a morphable face model to the measured depth data combining PSO and ICP techniques. Furthermore, [126] used a least-square technique to minimize the difference between the input depth change rate and the prediction rate, to perform 3D head tracking. Finally, in [241] a generative model is proposed, that unifies pose tracking and face model adaptation on-the-fly.
However, no previous method that uses depth maps as the only input exploits CNNs in an effective way. In this work we propose a method based on [20] which uses depth maps to produce accurate head pose predictions by leveraging CNNs.

RGB-D. RGB-D methods combine together RGB images and depth maps. A first effort to leverage both data was done in [239], where a Neural Network is exploited to perform head pose predictions. HOG features [45] were extracted from RGB and depth images in [279, 232], then a *Multi Layer Perceptron* and a linear SVM [96] were used for feature classification, respectively. In [119] Random Forests and tensor regression algorithms are exploited while [260] used a cascade of tree classifiers to tackle extreme head pose estimation task. Recently, in [181] a multimodal CNN was proposed to estimate gaze direction: a regression approach was only approximated through a classifier with a granularity of 1° and with 360 classes. As for RGB and depth methods, these appearance-based techniques are not robust enough: they still strongly depend on the detection of visible facial features.
Following 3D model registration techniques, [11] leverage intensity and depth data to build a 3D constrained local method for robust facial feature tracking. Furthermore, in [75, 27, 16, 142] a 3D morphable model is fitted, using both RGB and depth data to predict head pose. Finally, [216], based on a particle filter formalism, presents a new method for 3D face pose tracking in color images and depth data acquired by RGB-D cameras. Several works based on head pose estimation, however, do not take in consideration the head localization task.

Attributes Recognition

Early works on attribute recognition usually treat attributes independently training a different classifier for each attribute. Those methods involve the use of AdaBoost, K Nearest Neighbors [306] or SVM [51]. More recently Convolutional Neural Networks, enable researchers to mine the relationship between attributes and are preferred on large scale object recognition problems because of their advanced performances. There are large bodies of work on CNNs, like [116] which undertake the task of occlusion and low-resolution robust facial gender classification, or [88, 301] that predict facial attributes from faces in the wild. Many other works like [137, 294] propose different methods to achieve attribute classification like gender, smile and age in an unconstrained environment. However, those technique involve only facial images and are not suitable for surveillance tasks. Moreover [36, 131] address respectively the problem of describing people based on clothing attribute and the problem of clothing identification. Nevertheless, our work encompass the person as a whole and does not focus only on clothing classification.

More recent works that rely on full-body images to infer human attributes are

the Attribute Convolutional Net (ACN) and Deep learning based Multi-Attribute joint Recognition model (DeepMAR), [249, 139]. ACN jointly learns different attributes through a jointly-trained holistic CNN model, while DeepMAR utilizes the prior knowledge in the object topology for attribute recognition. In [295, 79, 80] attributes classification is accomplished combining part-based models and deep learning by training pose-normalized CNNs.

Additionally, MLCNN [307] splits the human body in 15 parts and train a CNN for each of them while DeepMAR* [141] divides the whole body in three parts which correspond to the headshoulder part, upper body and lower body of a pedestrian respectively. Furthermore, [144] tackles the problem of attribute recognition improving a part-based method within a deep hierarchical context.

Nevertheless, the majority of those methods relies on high resolution images and does not encompass the problem of occlusion. There are previous works that attempt to solve the problem of occlusion: [52] and [53] both leverage a large image database to find similar faces in order to complete the missing patch, but results are only shown for low resolution grey scale images.

2.3 Domain Translation

Domain translation is the task of learning a parametric translation function between two domains. Generative image modeling with deep learning techniques has received lots of attention in recent years. With the goal of learning a mapping from input to output images, works on this field can be split into two categories: unsupervised and supervised approaches.

Unsupervised

The first line of works follows the unsupervised setup. Here, the variational autoencoders (VAE) proposed by [218] and [125] are the first popular methods which apply a re-parameterization trick to maximize the lower bound of the data likelihood. The most popular methods are indeed generative adversarial networks (GAN) of [83] and [210], which simultaneously learn a generator network to generate image samples, and a discriminator network to discriminate generated samples from real ones. GANs are capable of generating sharp images by exploiting the adversarial loss instead of more canonical losses such as L1 or L2. Among the most successful methods, Isola *et al.* [107] demonstrated that their model, namely *pix2pix*, is effective at synthesizing photos from label maps, reconstructing objects

from edge maps and colorizing images. Moreover, Wang *et al.* [268] proposed a method that acts as a rendering engine: given a synthetic scene, their *Style GAN* is able to render a realistic image. In [297] a cGAN is capable of translating an RGB face image to depth data. Recently, a coupled generative adversarial networks framework has been proposed [150], to generate pairs of corresponding images in two different domains. In our preliminary work [20], we proposed one of the first approach, based on a traditional CNN with common aspects with respect to autoencoders [167] and Fully Convolutional Networks [155], that was trained to compute the appearance of a face using the corresponding depth information.

Supervised

The second group of works produces images conditioned on either categories, attributes, labels, images or texts. [277] proposed a Conditional Variational Autoencoder (CVAE) to achieve an image generation conditioned on attributes. On the other hand, [177] proposed conditional GANs (CGAN) where both the generator and the discriminator are conditioned on extra information to perform category specific image generation. [132] generated people in clothing, by conditioning on the fine-grained body part segments. [213] proposed a novel deep architecture and GAN formulation to effectively translating visual concepts from characters to pixels, by adding textual information to both generator and discriminator. They also further investigated the use of additional location, key-points, or segmentation information, to generate images as did by [215, 214]. With only these visual hints as condition and in contrast to our explicit condition on the occluded image, the control exerted over the image generation procedure is still abstract. Many works perform a conditioning over image generation not only on labels or texts but also on images. [299] generated multi-view cloth images from only a single view input by proposing a new image generation model that combines the strengths of the variational inference and the GAN framework. [33] tackled the unseen view inference by casting the problem in terms of tensor completion and adopt a factorization approach to accommodate single-view images. [108] provides a general purpose architecture that is effective at synthesizing photos from label maps, reconstructing objects from edge maps, and colorizing images, among other tasks. [280], [101], [286], [76] addressed the task of face image generation conditioned on a reference image and a specific face viewpoint. Finally, [278, 285, 202, 264] tackled the task of image inpainting where large missing regions have to be filled based on the available visual data. Our work can be seen as a particular case of inpainting, where the portion of the image to inpaint is not known a priori.

Chapter 3

Human Behaviour Understanding

In this chapter we discuss the core of the research done in these past three years, where synthetic data plays an important role in all the topics related to human behaviour understanding. In fact, the newly generated datasets enable a series of possibilities that go from new challenging tasks to more strict benchmark evaluations. In particular, in Section 3.1 synthetic data unlocks a new ambitious task, namely joint occlusion detection, where joint locations must be predicted even under severe occlusion. In Section 3.2 a new solution is enabled for tackling multi-person 3D human pose estimation in a bottom up fashion. Thanks to the newly generated data, we are able to evaluate a precise 3D pose estimation for more than 60 people with constant running time regardless of the number of subjects in the image. In Section 3.3 the synthetic data let us evaluate a system that targets human safety with a challenging benchmark, increasing his robustness. In Section 3.4 we leverage the generative power of GANs for hallucinating the occluded portion of a person relying in ad hoc generated data for the purpose. No one of the aforementioned tasks would have been feasible without exploiting synthetic data, either for the lack of available public data or for the impossibility of manual annotation and data collection.

3.1 2D Pose Estimation and Tracking

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 [175, 4]. 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 [247, 87, 46]. Nevertheless, in the last years several researchers [69, 247] 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 [247] 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 [29] 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;

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 [175, 4] demonstrate that the model can transfer positively towards real scenarios.

3.1.1 Dataset

The most widely used publicly available datasets for human pose estimation in videos are presented in Tab. 3.1. [296, 111, 32] provide annotations for the single-

Table 3.1: 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 [296]	2,326	159,633	1				✓	sports
JHMDB [111]	5,100	31,838	1				✓	diverse
YouTube Pose [32]	50	5,000	1				✓	diverse
Video Pose 2.0 [235]	44	1,286	1				✓	diverse
Posetrack [4]	514	23,000	1-13			✓	✓	diverse
MOT-16 [176]	14	11,235	6-51		✓	✓		urban
JTA	512	460,800	0-60	✓	✓	✓	✓	urban

person subtask of person pose estimation. Only Posetrack [4] 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 [176] 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.

We collected a massive dataset JTA (Joint Track Auto) 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 3.1. 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 3.1.

In order to virtually re-create real-world scenarios we manually directed the scenes by developing a game modification that interacts synchronously with the video game's engine. The developed module allowed us to generate and record

Figure 3.1: Examples from the JTA dataset exhibiting its variety in viewpoints, number of people and scenarios. Ground truth joints are superimposed to the original images. See supplementary material for further examples

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 almost 10M annotated body poses over 460,800 densely annotated frames. The distance from the camera ranges between 0.1 and 100 meters, resulting in pedestrian heights between 20 and 1100 pixels (see supplementary material for further details).

We collected a set of 512 Full 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 [3] and [4] 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 [4], which uses the annotated head bounding boxes as an estimation of the absolute scale of the

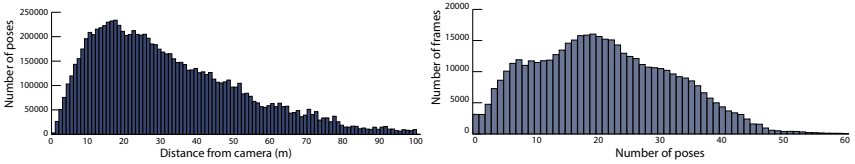

Figure 3.2: Statistics about JTA Dataset. (a) Number of annotated pose per camera distance (in meters). (b) Number of frames vs. the number of annotated poses per frame

person, we provide the precise scale of each pedestrian through the 3D annotation.

Figure 3.2 shows information about camera distances and poses per frame of the JTA Dataset. Figure 3.3 and Figure 3.4 provide examples from the dataset exhibiting its variety in viewpoints, number of people, illuminations and scenarios. The dataset, along with the game modification, are freely accessible at `http://imagelab.ing.unimore.it/jta`.

3.1.2 Method

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 [29] 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 3.5, 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 3.6, 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 the previous frame.

Single Image Pose Prediction

Our single image model, Figure 3.5, consists of an initial feature extractor based on the first 10 layers of VGG-19 [245] pretrained on COCO 2016 keypoints dataset [149]. 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 the heatmaps of the visible parts, the second branch predicts the heatmaps of the occluded parts and the third branch predicts

Figure 3.3: Some images taken from JTA Dataset (I)

Figure 3.4: Some images taken from JTA Dataset (II)

the part affinity fields (PAFs), which are vector fields used to link parts together. Note that, oppositely to [29], 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:

$$l_X^s = \sum_i \sum_{x=1}^{w'} \sum_{y=1}^{h'} M(x,y) \odot (X_i^s(x,y) - X_i^*(x,y))^2, \tag{3.1}$$

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 $L = \sum_{s=1}^{S} (l_H^s + l_O^s + l_P^s)$.

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 3.6. The temporal model takes as input N RGB frames of size $w \times h$ and produces,

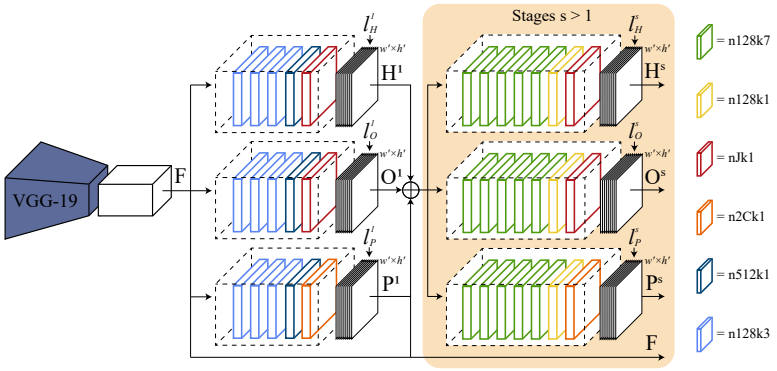

Figure 3.5: Architecture of the three-branch multi-stage CNN with corresponding kernel size (k) and number of feature maps (n) indicated for each convolutional layer

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 3.1.2, 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 3.6. As in Section 3.1.2, 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

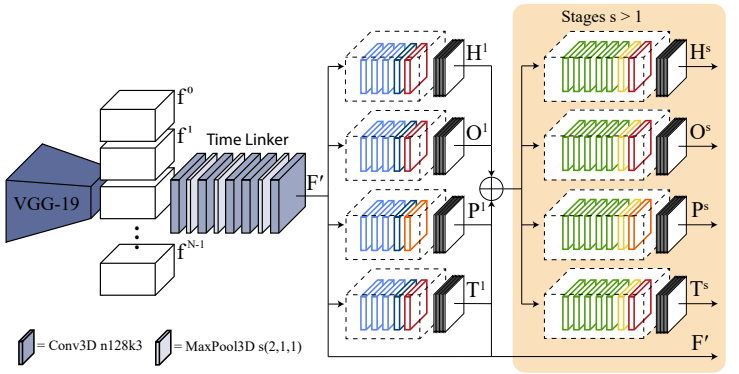

Figure 3.6: Architecture of our method that encompass pose estimation and tracking in an end-to-end fashion. The MaxPool3D perform pooling operations only in the temporal dimension with stride s

(visible and occluded) and PAFs computation. The complete network objective function then becomes $L = \sum_{s=1}^{S}(l_H^s + l_O^s + l_P^s + l_T^s)$ where

$$l_T^s = \sum_{j=1}^{J} \sum_{x=1}^{w'} \sum_{y=1}^{h'} M(x,y) \odot (T_j^s(x,y) - T_j^*(x,y))^2 \tag{3.2}$$

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 $T_j \in \mathbb{R}^{w \times h}, j \in \{1, \ldots, J\}$.

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 [29], 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 $p \in \mathbb{R}^2$

Figure 3.7: (a) Visualization of TAFs for different parts: for clarity, we show a single joint TAF for each person where color encodes direction. (b) Pose prediction performed on JTA dataset which distinguish between visible and occluded joints

results:

$$H_j^*(p) = \max_k \exp\left(-\frac{\left\|p - q_{j,k}\right\|_2^2}{\sigma^2}\right), \qquad \sigma = \exp\left(1 - \frac{d}{\alpha}\right) \tag{3.3}$$

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 3.7.(b)). 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 $q_{j,k}^{t-1}$ and $q_{j,k}^t$ 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.

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 $D_j = \{d_j^m : \text{for } j \in \{1, \dots, J\}, m \in \{1, \dots, N_j\}\}$ 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 $z_{j_1 j_2}^{mn} \in \{0, 1\}$ to indicate whether two joints candidates $d_{j_1}^m$ and $d_{j_2}^n$ are connected. Consequently, the objective is to find the optimal assignment for the set of possible connections, $Z = \{z_{j_1 j_2}^{mn} : \text{for } j_1, j_2 \in \{1, \dots, J\}, m \in \{1, \dots, N_{j_1}\}, n \in \{1, \dots, N_{j_2}\}\}$. 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:

$$E(d_{j_1}, d_{j_2}) = \int_{u=0}^{u=1} PAF(p(u)) \cdot \frac{d_{j_2} - d_{j_1}}{\left\| d_{j_2} - d_{j_1} \right\|_2} du \tag{3.4}$$

$$G(d_j, \hat{d}_j) = \int_{u=0}^{u=1} TAF(t(u)) \cdot \frac{\hat{d}_j - d_j}{\left\| \hat{d}_j - d_j \right\|_2} du \tag{3.5}$$

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):

$$\max_{Z_c} E_c = \max_{Z_c} \sum_{m \in D_{j_1}} \sum_{n \in D_{j_2}} (E(d_{j_1}^m, d_{j_2}^n) + \alpha E(\hat{d}_{j_1}^m, \hat{d}_{j_2}^n)) \cdot z_{j_1 j_2}^{mn}, \tag{3.6}$$

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

$$\hat{d}_{j_1}^m = \arg\max_{\hat{d}_{j_1}^b} G(d_{j_1}^m, \hat{d}_{j_1}^b), \qquad \hat{d}_{j_2}^n = \arg\max_{\hat{d}_{j_2}^q} G(d_{j_2}^n, \hat{d}_{j_2}^q) \tag{3.7}$$

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 (3.6) mixes both the contribution coming from the PAF in

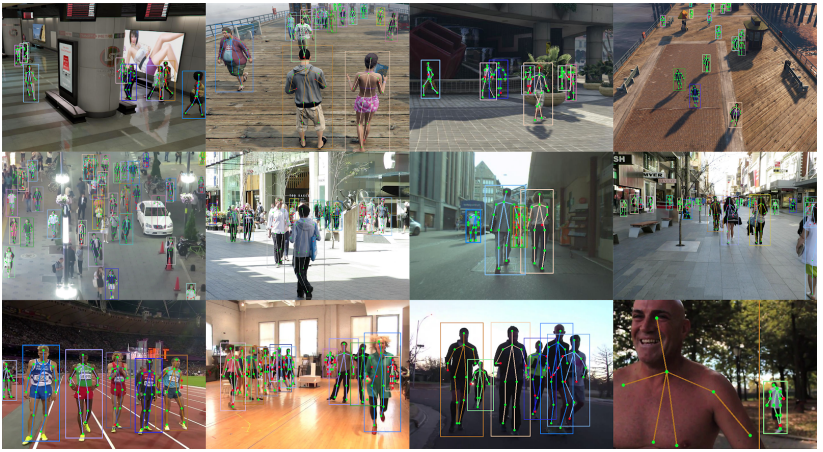

Figure 3.8: Qualitative results of THOPA-net on JTA (top row), MOT-16 (middle row) and PoseTrack (bottom row)

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 (3.6) 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, $E = \sum_{c=1}^{C} \max_{Z_c} E_c$.

3.1.3 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

Table 3.2: 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**
[29]	50.1	86.3	55.8	69.5

into account when extracting TAFs, thus exploited only for very short-term target association (namely tracklet construction).

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 [106], 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 3.2 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

Table 3.3: 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**
[29]	62.5	54.3	47.5	40.6	54.0	45.5	37.9	50.1

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 3.7.(b). 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 3.7.(a). Additionally we compared with [29] 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 [29] is the same as our *Single Image no occ* model in Table 3.2, with the only difference that the latter has been trained with distance rescaled versions of heatmaps and PAFs, according to Section 3.1.2, and it deals with multiple scales without any input rescaling operation. Table 3.3 also report per joint results in term of mean average precision .

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 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

Table 3.4: Tracking Results on JTA Dataset

	MOTA	IDF1	MT	ML	FP	FN	IDs	FRAG
[247] + our det	57.4	57.3	45.3	21.7	40096	103831	15236	15569
[247] + 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**

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), [247], inputed with either our detections and DPM [70] ones. Table 3.4 reports results in terms of Clear MOT tracking metrics [175]. 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.

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 two datasets: the commonly used MOT-16 Challenge Benchmark [175] and the new PoseTrack Dataset [4].

MOT-16. The MOT-16 Challenge Benchmark 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 3.8. All results are expressed in terms of Clear MOT metrics according to the benchmark protocol [175] and as for the virtual world tracking experiment the tracks were associated by maximizing the TAF scores between detections. The network was end-to-end fine-tuned, with the exception of

Table 3.5: Results on MOT-16 benchmark ranked by MOTA score

	MOTA	IDF1	MT	ML	FP	FN	IDs	FRAG
[287]	**66.1**	**65.1**	**34.0**	20.8	5061	**55914**	805	3093
[272]	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
[231]	47.2	46.3	14.0	41.6	**2681**	92856	774	1675
[40]	46.0	50.0	14.6	43.6	6895	91117	**473**	**1422**
[8]	43.9	45.1	10.7	44.4	6450	95175	676	1795
[234]	38.8	42.4	7.9	49.1	8114	102452	965	1657

Table 3.6: 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

the occlusion branch. 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.

Table 3.5 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. 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 and FRAG given by the fact that our output is an intermediate step between detections and long-term tracking. Nevertheless, we remark that

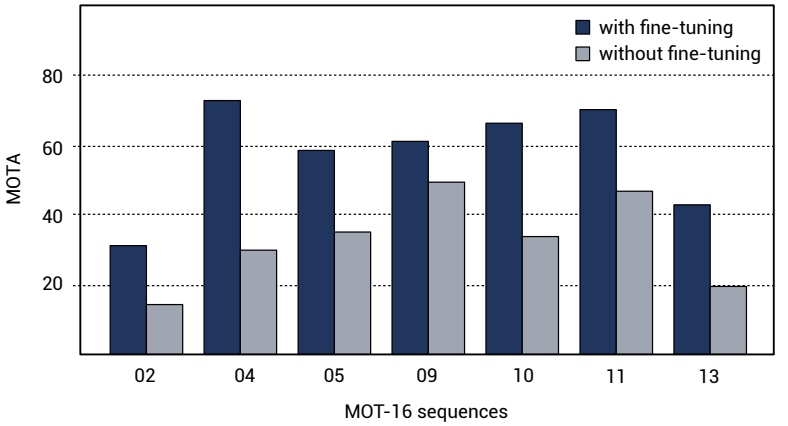

Figure 3.9: Bar-chart of THOPA-net (trained on JTA dataset) results on MOT16 training set with and without fine-tuning on real data

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 [287] and [272] 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 [231] 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 3.8 shows qualitative results of our proposal.

Table 3.6 reports the metrics per sequence performed on MOT-16. A further experiment was conducted on the MOT-16 benchmark training set Figure 3.9 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

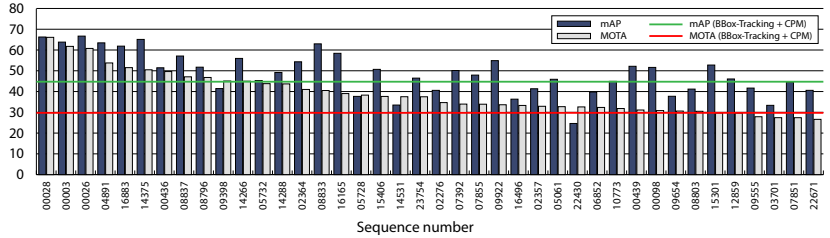

Figure 3.10: Results on PoseTrack dataset compared with a BBox-Tracking + CPM (trained on MPII) baseline (used also in PoseTrack; red/green lines are the average of performances on the selected sequences to avoid plot clutter)

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.

PoseTrack. The PoseTrack Dataset is a large-scale benchmark for multi-person pose estimation and tracking in videos. It contains 550 videos including around 23,000 annotated frames, split into 292, 50, 208 videos for training, validation and testing, respectively. The annotations include 15 body keypoints location, a unique person id and a head bounding box for each person instance. We tested our solution on a subset of PoseTrack Dataset with surveillance like features (e.g. people standing, walking, etc.). We remark that PoseTack exhibits different features w.r.t. surveillance context in which the targets number is higher and the camera FoV is mostly a far FoV. In Fig. 3.10 we show MOTA and mAP results of THOPA-net on PoseTrack sequences (solely using synthetic data for training). We used training and validation sequences in order to obtain per-sequence results. The results are satisfying (see Fig 3.8) 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.

3.2 3D Pose Estimation

Human Pose Estimation (HPE) has seen significant progress in recent years, mainly thanks to deep Convolutional Neural Networks (CNNs). Best performing methods on 2D HPE are all leveraging heatmaps to predict body joint locations [29, 274, 253]. Heatmaps have also been extended for 3D HPE, showing promising results in single person contexts [236, 205, 250].

Despite their good performance, these methods do not easily generalize to multi-person 3D HPE, mainly because of their high demands for memory and computation. This drawback also limits the resolution of those maps, that have to be kept small, leading to quantization errors. Using larger volumetric heatmaps can address those issues, but at the cost of extra storage, computation and training complexity.

In this thesis, we propose a simple solution to the aforementioned problems that allows us to directly predict high-resolution volumetric heatmaps while keeping storage and computation small. This new solution enables our method to tackle multi-person 3D HPE using heatmaps in a single-shot bottom-up fashion. Moreover, thanks to our high-resolution output, we are able to produce fine-grained absolute 3D predictions even in single person contexts. This allows our method to achieve state of the art performance on the most popular single person benchmark [104].

The core of our 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. We named this solution LoCO, *Learning on Compressed Output*.

By leveraging on the analogy between compression and dimensionality reduction on sparse signals [265, 238, 6], 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. our proposal, compression-based approaches have been effectively used for both dataset distillation and input compression [267, 257] 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.

Specifically, our HPE pipeline consists of two modules: at first, the pretrained

Volumetric Heatmap Autoencoder is used to obtain a smaller/denser representation of the volumetric heatmaps. These "codes" are then used to supervise the *Code Predictor*, which aims at estimating multiple 3D joint locations from a monocular RGB input.

To summarize, the novel aspects of our proposal are:

- We propose a simple and effective method that maps high-resolution volumetric heatmaps to a compact and more tractable representation. This saves memory and computational resources while keeping most of the informative content.
- This new data representation enables the adoption of volumetric heatmaps to tackle multi-person 3D HPE in a bottom-up fashion, an otherwise intractable problem. Experiments on both real [114] and simulated environments [63] (see Fig. 3.11) show promising results even in 100 meters wide scenes with more than 50 people. Our method only requires a single forward pass and can be applied with constant running time regardless of the number of subjects in the scene.
- We further demonstrate the generalization capabilities of LoCO by applying it to a single person context. Our fine-grained predictions establish a new state of the art on Human3.6m [104] among bottom-up methods.

3.2.1 Method

The following subsections summarize the key elements of LoCO. Section 3.2.1 gives a preliminary definition of the chosen volumetric heatmap representation and elaborates on its merits. Section 3.2.1 illustrates our proposed data mapping which addresses the high dimensional nature of the volumetric heatmaps by producing a compact and more tractable representation. Next, in Section 3.2.1, we describe how our strategy can be easily exploited to effectively tackle the problem of multi-person 3D HPE in a single-shot bottom-up fashion. Finally, Section 3.2.1 illustrates our simple refining approach that prevents poses from being implausible.

Volumetric Heatmaps

By considering a voxelization of the RGB-D volumetric space [44, 205], we refer as a volumetric heatmap, $\mathfrak{h}$, the 3D confidence map with size $D \times H \times W$, where D represents the depth dimension (appropriately quantized), while H and W

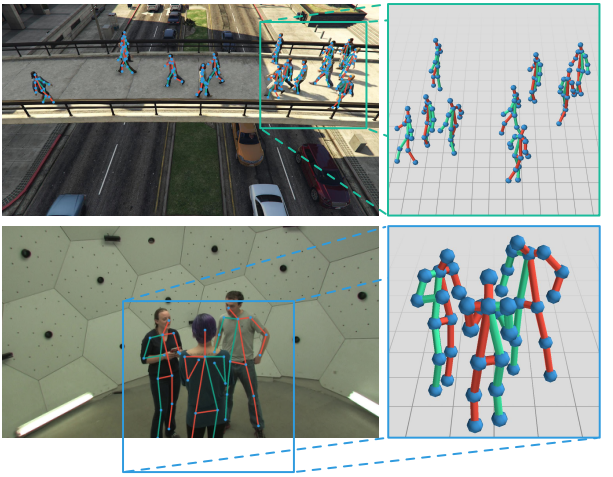

Figure 3.11: Examples of 3D poses estimated by our LoCO approach. Close-ups show that 3D poses are correctly computed even in very complex and articulated scenarios

represent the height and width of the image plane respectively. Given the body joint j with pseudo-3D coordinates $\mathbf{u}_j = (u_{1,j}, u_{2,j}, u_{3,j})$, where $u_{1,j} \in \{1, ..., D\}$ is the quantized distance of joint j from the camera, and $u_{2,j} \in \{1, ..., H\}$ and $u_{3,j} \in \{1, ..., W\}$ are respectively the row and column indexes of its pixel on the image plane, the value of $\mathfrak{h}_\mathrm{j}$ at a generic location u is obtained by centering a fixed variance Gaussian in u_j:

$$\mathfrak{h}_\mathrm{j}(\mathbf{u}) = e^{-\frac{\|\mathbf{u}-\mathbf{u}_j\|^2}{\sigma^2}} \tag{3.8}$$

In a multi-person context, in the same image we can simultaneously have several joints of the same kind (e.g. “left ankle"), one for each of the K different people in the image. In this case we aggregate those K volumetric heatmaps ${\mathfrak{h}_j}^{(k)}$, into a single heatmap $\mathfrak{h}_j$ with a max operation:

$$\mathfrak{h}_j(\mathbf{u}) = \max_k\{{\mathfrak{h}_j}^{(k)}(\mathbf{u})\} \tag{3.9}$$

Finally, considering N different types of joint and K people, we have a set of N volumetric heatmaps (each associated with a joint type), $\mathfrak{H} = \{\mathfrak{h}_j, j = 1, ..., N\}$,

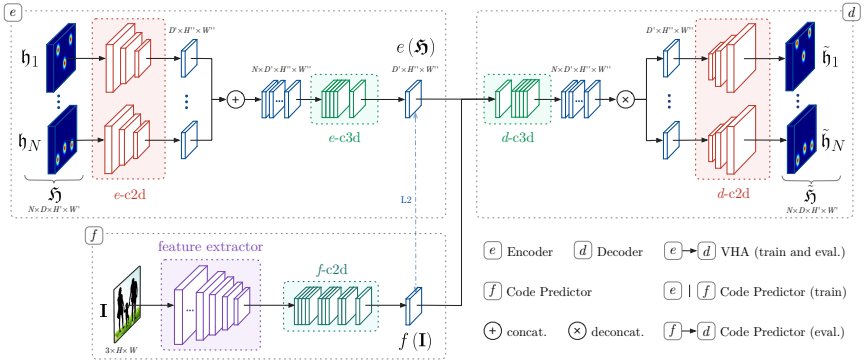

Figure 3.12: Schematization of the proposed LoCO pipeline. At training time, the Encoder e produces the compressed volumetric heatmaps $e(\mathfrak{H})$ which are used as ground truth from the Code Predictor f. At test time, the intermediate representation $f(I)$ computed by the Code Predictor is fed to the Decoder d for the final output. In our case, $H' = H/8$ and $W' = W/8$

resulting from the aggregation of the individual heatmaps of the K people in the scene. Note that, given pseudo-3D coordinates $\mathbf{u} = (u_1, u_2, u_3)$ and the camera intrinsic parameters, i.e. focal length $f = (f_x, f_y)$ and principal point (c_x, c_y), the corresponding 3D coordinates $\mathbf{x} = (x, y, z)$ in the camera reference system can be retrieved by directly applying the equations of the pinhole camera model.

The benefit of choosing a volumetric heatmap representation over a direct 3D coordinate regression is that it casts the highly non-linear problem to a more tractable configuration of prediction in a discretized space. In fact, joint predictions do not estimate a unique location but rather a per voxel confidence, which makes it easier for a network to learn the target function [205]. In the context of 2D HPE, the benefits of predicting confidences for each pixel instead of image coordinates are well known [207, 256]. Moreover, in a multi-person environment, directly regressing the joint coordinates is unfeasible when the number of people is not known a priori, making volumetric heatmaps a natural choice for tackling bottom-up multi-person 3D HPE.

The major disadvantage of this representation is that it is memory and computational demanding, requiring some compromise during implementation that limits its full potential. Some of those compromises consist in utilizing low resolution heatmaps that introduce quantization errors or complex training strategies that

block	layer	in ch.	out ch.	stride
e-c2d	Conv2D + ReLU	D	D/d_1	s_1
	Conv2D + ReLU	D/d_1	D/d_2	s_2
	Conv2D + ReLU	D/d_2	D/d_3	s_2
e-c3d	Conv3D + ReLU	N	4	1
	Conv3D + ReLU	4	1	1

Table 3.7: Structure of the encoder part of the Volumetric Heatmap Autoencoder (VHA). The decoder is not shown as it is perfectly mirrored to the encoder. *VHAv1*: $(d_1, d_2, d_3) = (1, 2, 2)$ and $(s_1, s_2, s_3) = (1, 2, 1)$; for *VHAv2*: $(d_1, d_2, d_3) = (2, 4, 4)$ and $(s_1, s_2, s_3) = (2, 2, 1)$; *VHAv3*: $(d_1, d_2, d_3) = (2, 4, 8)$ and $(s_1, s_2, s_3) = (2, 2, 2)$

involve coarse-to-fine predictions through iterative refining of network output [205].

Volumetric Heatmap Autoencoder

To overcome the aforementioned limitations without introducing quantization errors or training complexity, we propose to map volumetric heatmaps to a more tractable representation. Inspired by [157], we propose a multiple branches Volumetric Heatmap Autoencoder (VHA) that takes a set of N volumetric heatmaps $\mathfrak{H}$ as input. At first, the volumetric heatmaps $\{\mathfrak{h}_1, ..., \mathfrak{h}_N\}$ are processed independently with a 2D convolutional block (*e*-c2d) in which the kernel does not move along the D dimension. In order to capture the mutual influence between joints locations, the obtained maps are then stacked along a fourth dimension and processed by a subsequent set of 3D convolutions (*e*-c3d). The resulting encoded representation, $e(\mathfrak{H})$ is finally decoded by its mirrored architecture $d\left(e\left(\mathfrak{H}\right)\right) = \tilde{\mathfrak{H}}$. The general structure of the model is outlined in Fig. 3.12 top.

The goal of the VHA is therefore to learn a compressed representation of the input volumetric heatmaps that preserve their information content, which results in the preservation of the position of the Gaussian peaks of the various joints in the original maps. For the purpose, we maximize the F1-score, F1 $\left(Q_{\mathfrak{H}}, Q_{\tilde{\mathfrak{H}}}\right)$, between the set of ground truth peaks ($Q_{\mathfrak{H}}$) and the set of the decoded maps ($Q_{\tilde{\mathfrak{H}}}$). We define the set of peaks as follows:

$$Q_{\mathfrak{H}} = \bigcup_{n=1,\dots,N} \{\mathbf{u} : \mathfrak{h}_n(\mathbf{u}) > \mathbf{u}' \ \forall \mathbf{u}' \in \mathfrak{N}_{\bar{\mathbf{u}}}\} \tag{3.10}$$

where $\mathfrak{N}_{\bar{\mathbf{u}}}$ is the 6-connected neighborhood of $\bar{\mathbf{u}}$, i.e. the set of coordinates $\mathfrak{N}_{\bar{\mathbf{u}}} = \{\mathbf{u} : \|\mathbf{u} - \bar{\mathbf{u}}\| = 1\}$ at unit distance from $\bar{\mathbf{u}}$. Since the procedure for extracting the coordinate sets from the volumetric heatmaps is not differentiable, the former objective cannot be directly optimized as a loss component for training the VHA. To address this issue, we propose to use mean squared error (MSE) loss between $\mathfrak{H}$ and $\tilde{\mathfrak{H}}$ as training loss.

Note that our proposed mapping purposely reduces the volumetric heatmap's fourth dimension, making its shape coherent with the output of 2D convolutions and thus exploitable by regular CNN backbones. Additional architecture details can be found in the supplementary material.

Code Predictor and Body Joints Association

The input of the Code Predictor is represented by a RGB image, $\mathbf{I}$, while its output, $f(\mathbf{I})$, aims to predict the codes obtained with the VHA, Fig. 3.12. The architecture, Fig. 3.12 bottom, is inspired by [274] thus composed by a pre-trained feature extractor (convolutional part of Inception v3 [252]), and a fully convolutional block (f-c2d) composed of four convolutions. We trained the Code Predictor by minimizing the MSE loss between $f(\mathbf{I})$ and $e(\mathfrak{H})$, where $\mathfrak{H}$ is the volumetric heatmap associated with the image $\mathbf{I}$.

At inference time, the pseudo-3D coordinates of the body joints are obtained from the decoded volumetric heatmap $\tilde{\mathfrak{H}} = d(f(\mathbf{I}))$ through a local maxima search. Eventually, if camera parameters are available, the pinhole camera equations recover the true three-dimensional coordinates of the detected joints. Additional details in the supplementary material.

As in almost all recent 2D HPE bottom-up approaches [29, 71, 39] (i.e. methods which does not require a people detection step) detected joints have to be linked together to obtain people skeletal representations. In a single person context, joint association is trivial. On the other hand, in a multi-person environment, linking joints is significantly more challenging. For the purpose, we rely on a simple distance-based heuristic where, starting from detected heads (i.e. the joint with the highest confidence), we connect the remaining $(N-1)$ joints by selecting the closest ones in terms of 3D Euclidean distance. Associations are further refined by rejecting those that violates anatomical constraints (e.g. length of a limb greater than a certain threshold). Despite its simplicity, this approach is

		F1 on JTA		
model	bottleneck size	@0vx	@1vx	@2vx
VHA$^{(1)}$	$\frac{D}{2} \times \frac{H'}{2} \times \frac{W'}{2}$	97.1	98.4	98.5
VHA$^{(2)}$	$\frac{D}{4} \times \frac{H'}{4} \times \frac{W'}{4}$	92.5	97.0	97.1
VHA$^{(3)}$	$\frac{D}{8} \times \frac{H'}{8} \times \frac{W'}{8}$	56.5	90.3	92.9

		F1 on Panoptic		
model	bottleneck size	@0vx	@1vx	@2vx
VHA$^{(1)}$	$\frac{D}{2} \times \frac{H'}{2} \times \frac{W'}{2}$	-	-	-
VHA$^{(2)}$	$\frac{D}{4} \times \frac{H'}{4} \times \frac{W'}{4}$	97.1	98.6	98.9
VHA$^{(3)}$	$\frac{D}{8} \times \frac{H'}{8} \times \frac{W'}{8}$	91.9	98.7	99.6

		F1 on Human3.6m		
model	bottleneck size	@0vx	@1vx	@2vx
VHA$^{(1)}$	$\frac{D}{2} \times \frac{H'}{2} \times \frac{W'}{2}$	-	-	-
VHA$^{(2)}$	$\frac{D}{4} \times \frac{H'}{4} \times \frac{W'}{4}$	100.0	100.0	100.0
VHA$^{(3)}$	$\frac{D}{8} \times \frac{H'}{8} \times \frac{W'}{8}$	99.7	100.0	100.0

Table 3.8: VHA bottleneck/code size and performances on the JTA, Panoptic and Human3.6m (protocol P2) test set in terms of F1 score at different thresholds @0, @1, and @2 voxel(s); @*t* indicates that a predicted joint is considered "true positive" if the distance from the corresponding ground truth joint is less than *t*

particularity effective when 3D coordinates of body joints are available, especially in surveillance scenarios where proxemics dynamics often regulate the spatial relationships between different individuals. Additional details are reported in the supplementary material.

Pose Refiner

The predicted 3D poses are subsequently refined by a MLP network trained to account for miss-detections and location errors. The objective of the Pose Refiner is indeed to make sure that the detected poses are complete (i.e. all the N joints

	PR	RE	F1	PR	RE	F1	PR	RE	F1
	@0.4 m			@0.8 m			@1.2 m		
LM [172, 173]	5.8	5.3	5.4	24.0	21.6	22.2	41.4	36.9	38.2
LM [172, 173] + ref.	5.8	5.8	5.7	23.2	23.5	23.0	38.8	39.1	38.4
[212] + [166]	**75.8**	28.3	39.1	**92.8**	34.1	47.3	**96.3**	35.3	49.0
Uncompr. VH	25.3	24.4	24.4	45.4	43.1	43.5	55.5	52.4	53.0
LoCO$^{(1)}$	48.1	42.7	44.7	65.6	58.5	61.2	72.4	64.8	67.7
LoCO$^{(1)}$+.	49.3	43.4	45.7	66.8	59.0	62.0	73.5	65.0	68.2
LoCO$^{(2)}$	54.7	46.9	50.1	70.6	60.4	64.6	77.0	65.9	70.4
LoCO$^{(2)}$+.	55.3	**47.8**	**50.8**	70.6	**60.9**	**64.7**	76.8	**66.3**	**70.4**
LoCO$^{(3)}$	48.1	41.9	44.4	66.9	58.2	61.7	74.4	64.7	68.6
LoCO$^{(3)}$+.	49.1	42.8	45.3	67.1	58.4	61.9	74.3	64.7	68.5
GT LM	76.0	64.8	69.5	76.0	64.8	69.5	76.0	64.8	69.5
GT VH	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9

Table 3.9: Comparison of our LoCO approach with other strong baselines and competitors on the JTA test set. In PR (precision), RE (recall) and F1, @t indicates that a predicted joint is considered "true positive" if the distance from the corresponding ground truth joint is less than t. Last two rows contain the upper bounds obtained using the ground truth location maps and volumetric heatmaps respectively

are always present). To better understand how the Pose Refiner works, we define the concept of *3D poses* and *root-relative poses*. Given a person k, its 3D pose is the set $\mathbf{p}^{(k)} = \left\{\mathbf{x}_n^{(k)},\ n = 1, ..., N\right\}$ of the 3D coordinates of its N joints. The corresponding root-relative pose is then given by:

$$\mathbf{p_{rr}}^{(k)} = \left\{\frac{\mathbf{x}_n^{(k)} - \mathbf{x}_1^{(k)}}{l_n},\ n = 2, ..., N\right\} \tag{3.11}$$

where $\mathbf{x}_1$ are the 3D coordinates of the root joint ("head-top" in our experiments) and l_n is a normalization constant computed on the training set as the maximum length of the vector that points from the root joint to any other joint of the same person.

The Pose Refiner is hence trained with MSE loss taking as input the root-relative version of the 3D poses with randomly removed joints, and an additional Gaussian noise applied to the coordinates. Given the 3D position of the root joint

and the refined poses, it is straightforward to re-obtain the corresponding 3D poses by using Eq. (3.11).

3.2.2 Experiments

A series of experiments have been conducted on two multi-person datasets, namely JTA [63] and CMU Panoptic [114, 244, 115], as well as one well established single-person benchmark: Human3.6m [104].

JTA is a large synthetic dataset for multi-person HPE and tracking in urban scenarios. It is composed of 512 Full HD videos, 30s long, each containing an average of 20 people per frame. Due to its recent publication date, this dataset does not have a public leaderboard and it is not mentioned in other comparable HPE works. Despite this limitation, we believe it is crucial to test LoCO on JTA because it is much more complex and challenging than older benchmarks.

CMU Panoptic is another large dataset containing both single-person and multi-person sequences for a total of 65 sequences (5.5 hours of video). It is less challenging than JTA as the number of people per frame is much more limited, but it is currently the largest real-world multi-person dataset with 3D annotations.

To further demonstrate the generalization capabilities of LoCO, we also provide a direct comparison with other HPE approaches on the single person task. Without any modification to the multi-person pipeline, we achieve state of the art results on the popular Human3.6m dataset.

For each dataset we also show the upper bound obtained by using the GT volumetric heatmaps in order to highlight the strengths of this data representation. In all the following tables, we will indicate with LoCO$^{(n)}$ our complete HPE pipeline, composed of the Code Predictor, the decoder of VHA$^{(n)}$ and the subsequent post-processing. LoCO$^{(n)}+$ is the same system with the addition of the Pose Refiner.

For all the experiments related to this work we utilized Adam optimizer with learning rate 10^{-4}. We employed batch size 1 when training the VHA and batch size 8 when training the Code Predictor. We employed Inception v3 [252] as backbone for the Code Predictor, which is followed by 3 convolutions with ReLU activation having kernel size 4 and with 1024, 512 and 256 channels respectively. A last 1×1 convolution is performed to match the compressed volumetric heatmap's number of channels. Additional training details in the supplementary material.

Compression Levels

In order to understand how different code sizes in the VHA affects the performance of our Code Predictor network, multiple VHA versions have been tested. Specifically, we designed three VHA versions with decreasing bottleneck sizes. Each version has been trained on JTA first and then finetuned on CMU Panoptic and Human3.6m. VHA's architecture details are depicted in Tab. 3.7 for every version.

As shown in Tab. 3.8, as the bottleneck size decreases, there is a corresponding decrease in the F1-score. Intuitively, the more we compress, the less information is being preserved. VHA$^{(1)}$ is only considered when using JTA, as VHA$^{(2)}$ and VHA$^{(3)}$ already obtain an almost lossless compression on Panoptic and Human3.6m, due to their smaller number of people in the scene.

All the experiments has been conducted considering a 14 joints volumetric heatmap representation of shape $14 \times D \times H' \times W'$, where H' and W' are height and width downsampled by a factor of 8, while D has been fixed to 316 bins. Note that the real-world depth grid covered by our representation is a uniform discretization in $[0, 100]$m for JTA, $[0, 7]$m for Panoptic and $[1.8, 8.1]$m for Human3.6m. Thus, every bin has a depth size of approximately 0.32m for JTA and 0.02m for Panoptic and Human3.6m.

HPE Experiments on JTA Dataset

On the JTA dataset we compared LoCO against the Location Maps based approaches of [172, 173]. Currently the Location Maps representation is the most relevant alternative to volumetric heatmaps to approach the 3D HPE task in a bottom-up fashion and therefore represents our main competitor.

A Location Maps is a per-joint feature channel that stores the 3D coordinate x, y, or z at the joint 2D pixel location. For each joint there are three location-maps and the 2D heatmap. The 2D heatmap encodes the pixel location of the joint as a confidence map in the image plane. The 3D position of a joint can then be obtained from its Location Map at the 2D pixel location of the joint. For a fair comparison, we utilized the same network (Inception v3 + f-c2d) to directly predict the Location Maps. The very low F1 score demonstrate that Location Maps are not suitable for images with multiple overlapping people, not being able to effectively handle the challenging situations peculiar of crowded surveillance scenarios (see Tab. 3.9).

Additionally, we report a comparison with a strong top-down baseline that

Figure 3.13: Qualitative results of LoCO$^{(2)}+$ on the JTA and Panoptic datasets. We show both the 3D poses (JTA: 2nd row, Panoptic: 4th row) and the corresponding 2D versions re-projected on the image plane (JTA: 1st row, Panoptic: 3rd row)

uses YOLOv3 [212] for the people detection part and [166] as the single-person pose estimator. [166], like almost all single person methods, provides root-relative joint coordinates and not the absolute 3D position. We thus performed the 3D alignment according to [226] by minimizing the distance between 2D pose and re projected 3D pose. We outperform this top-down pipeline by a large margin in terms of F1-score, while being significantly faster; LoCO is able to process Full HD images with more than 50 people at 8 FPS on a Tesla V100 GPU, while the top-down baseline runs at an average of 0.5 FPS (16 times slower). The recall gap is mostly due to the fact that the detection phase in top-down approaches usually miss overlapped or partially occluded people on the crowded JTA scenes.

Finally, we compared against an end-to-end model trained to directly predict the volumetric heatmaps without compression ("Uncompr. Volumetric Heatmaps" in Tab. 3.9). Specifically, we stacked the Code Predictor and the VHA$^{(2)}$'s decoder and trained it in an end-to-end fashion. Our technique outperforms this version at every compression rate. In fact, the sparseness of the target makes it difficult to effectively exploit the redundancy of body poses in the ground truth annotation

leading to a more complex training phase.

We point out that LoCO$^{(2)}+$ obtains by far the best result in terms of F1-score compared to all evaluated approaches and baselines, thus demonstrating the effectiveness of our method. Moreover, the best result has been obtained using the VHA$^{(2)}$'s mapping, which seemingly exhibits the best compromise between information preserved and density of representation. It is also very interesting to note that the upper bound for Volumetric Heatmaps is much higher than that of Location Maps (last two rows of Tab. 3.9), highlighting the superiority of volumetric heatmaps in crowded scenarios. It is finally worth noticing that LoCO$^{(1)}+$ and LoCO$^{(3)}+$ obtain very close results, indicating that an extremely lossy compression can lead to a poor solution as much as utilizing a too sparse and oversized representation.

Following the protocol in [63], we trained all our models (and those with Location Maps) on the 256 sequences of the JTA training set and tested our complete pipeline only on every 10$^{\text{th}}$ frame of the 128 test sequences. Qualitative results are presented in Fig. 3.13.

HPE Experiments on Panoptic Dataset

Here we propose a comparison between LoCO and three strong multi-person approaches [291, 290, 209] on CMU Panoptic following the test protocol defined in [290]. The results, shown in the Tab. 3.10, are divided by action type and are expressed in terms of Mean Per Joint Position Error (MPJPE). MPJPE is calculated by firstly associating predicted and ground truth poses, by means of a simple Hungarian algorithm. In the Tab. 3.10 we also report the F1-score: the solely MPJPE metric is not meaningful as it does not take into account missing detections or false positive predictions.

The obtained results show the advantages of using volumetric heatmaps for 3D HPE, as LoCO$^{(2)}+$ achieves the best result in terms of average MPJPE on the Panoptic test set. For the sake of fairness, we also tested on the no longer maintained "mafia" sequence. However, the older version of the dataset utilizes a different convention for the joint positions. This, in fact, is reflected by the worst performance in that sequence only. Once again, the best trade-off is obtained using VHA$^{(2)}$, due to VHA$^{(3)}$'s mapping partial loss of information. The GT upper bound in Tab. 3.10 further demonstrate the potential of our representation. Qualitative results are presented in Fig. 3.13.

	MPJPE [mm]					
	Haggl.	Mafia	Ultim.	Pizza	Mean	F1
[209]	218	187	194	221	203	-
[290]	140	166	151	156	153	-
[291]	72	**79**	67	94	72	-
LoCO$^{(2)}$+	**45**	95	**58**	**79**	**69**	89.21
LoCO$^{(3)}$+	48	105	63	91	77	87.87
GT	9	12	9	9	10	100

Table 3.10: 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

HPE Experiments on Human3.6m Dataset

In analogy with previous experiments, we tested LoCO on Human3.6m. Unlike most existing approaches, we apply our multi-person method as it is, without exploiting the knowledge of the single-person nature of the dataset, as we want to demonstrate its effectiveness even in this simpler context. Results, with and without rigid alignement, are reported in terms of MPJPE following the P1 and P2 protocols. In the P1 protocol, six subjects (S1, S5, S6, S7, S8 and S9) are used for training and every 64$^{\text{th}}$ frame of subject S11/camera 2 is used for testing. For the P2 protocol, all the frames from subjects S9 and S11 are used for testing and only S1, S5, S6, S7 and S8 are used for training.

Tab. 3.11 shows a comparison with recent state-of-the-art multi-person methods, showing that our method is well suited even in the single person context, as LoCO$^{(3)}$+ achieves state of the art results among bottom up methods. Note that, although Moon *et al.* reports better numerical performance, they leverage additional data for training and evaluate on a more redundant set of joints containing pelvis, torso and neck. It is worth noticing that LoCO$^{(3)}$+ performs substantially better than LoCO$^{(2)}$+, demonstrating that a smaller representation is preferred when the same amount of information is preserved (99.7 and 100.0 F1@0vx respectively on VHA$^{(3)}$ and VHA$^{(2)}$). Qualitative results are presented in Fig. 3.14.

	method	N	P1	P1 (a)	P2	P2 (a)
top-down	Rogez *et al.* [225]	13	63.2	53.4	87.7	71.6
	Dabral *et al.* [43]	16	-	-	-	65.2
	Rogez *et al.* [226]	13	54.6	45.8	65.4	54.3
	Moon *et al.* [178]	17	**35.2**	**34.0**	**54.4**	53.3
bottom-up	Mehta *et al.* [173]	17	-	-	80.5	-
	Mehta *et al.* [172]	17	-	-	69.9	-
	LoCO$^{(2)}+$	14	84.0	75.4	96.6	77.1
	LoCO$^{(3)}+$	14	51.1	43.4	61.0	**49.1**
GT Vol. Heatmaps		14	15.6	14.9	15.0	14.3

Table 3.11: 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. Last row: results with ground truth volumetric heatmaps

Detection Experiments

To show how our LoCO approach can be effectively adopted also for the detection task in crowded scenarios under heavy occlusions, we have tested our system in terms of 2D people detection comparing it with YOLOv3 [212] on the JTA test set. Using LoCO, we predict 3D poses and project them on 2D bounding boxes using the camera intrinsic parameters.

In terms of precision, recall, and F1 (with the bounding box IoU threshold at 0.5), using our LoCO$^{(2)}+$ trained on JTA, we get 81.94, 69.73, and 75.39 respectively; with out of the box YOLOv3, instead, we obtain 99.12, 30.81 and 44.50.

Although our model is less precise than YOLOv3 (around -20%), it surpasses it by a large margin (around +40%) in terms of recall, resulting in an F1-score that is clearly in our favor (almost +30%). The scenes in JTA, in fact, are extremely crowded and present a very high percentage of occlusion with multiple overlapping people. It is not easy for a detector to handle situations of this type, while a part-based bottom-up method is much less affected by this problem.

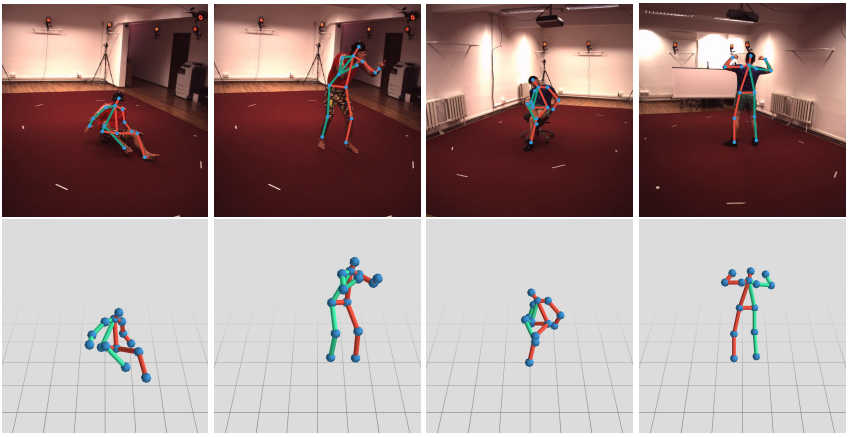

Figure 3.14: Qualitative results of LoCO$^{(3)}$+ on the Human3.6m dataset

3.2.3 Implementation

For the sake of reproducibility, in this section we illustrate the architectures of the Code Predictor and the Pose Refiner modules of our LoCO pipeline. Code and models available at `https://github.com/fabbrimatteo/LoCO`.

Code Predictor Inspired by [274], our method simply adds a few convolutional layers (f-c2d) to the last convolution stage of a backbone network. Tab. 3.12 reports the detailed structures of the various f-c2d blocks utilized in our experiments. *ConvTr2D* refers to transposed 2D convolutions while *Conv2D* refers to simple 2D convolutions. For each layer, we provide: number of input channels, number of output channels, kernel size and stride. In all the proposed experiments we utilized InceptionV3 [252] pretrained on Imagenet [49] as backbone architecture.

Pose Refiner The structure of the Pose Refiner is shown in Fig. 3.15. It is a simple network composed by three fully connected layers with ReLU activation followed by a skip connection. Input and output are normalized root-relative representations of a single 3D pose, with values in range $[0, 1]$. During training, Gaussian noise (mean: $0m$, variance: $0.08m$) is applied to the input pose while some joints are randomly removed with probability 0.1. The removed joints are

	layer	in ch.	out ch.	ker.	str.
$\text{LoCO}^{(1)}$	ConvTr2D→ReLU	F	1024	4	2
	ConvTr2D→ReLU	1024	512	4	2
	Conv2D→ReLU	512	256	4	1
	Conv2D	256	158	1	1
$\text{LoCO}^{(2)}$	ConvTr2D→ReLU	F	1024	4	2
	Conv2D→ReLU	1024	512	4	1
	Conv2D→ReLU	512	256	4	1
	Conv2D	256	79	1	1
$\text{LoCO}^{(3)}$	Conv2D→ReLU	F	1024	4	1
	Conv2D→ReLU	1024	512	4	1
	Conv2D→ReLU	512	256	4	1
	Conv2D	256	39	1	1

Table 3.12: Structure of the three f-c2d block variants of the Code Predictor used for our HPE experiments. F represents the number of output channels of the exploited feature extractor. In all our experiments we used Inception v3 with $F = 2048$

coded with a default value of $(-1, -1, -1)$.

Volumetric Heatmap Spaces In our experiments, we defined our Volumetric Heatmap representation according to two different pseudo-3D spaces, depending on which dataset we used:

- The first space is defined as $S_1 = D \times H' \times W'$, where H' and W' are the height and width, downsampled by a actor of 8, of the image plane and D is the maximum distance from the camera in meters, quantized with 316 bins. We adopted S_1 for JTA.

- The second space is defined as $S_2 = Z \times H' \times W'$, where H' and W' are defined as in S_1, and Z is the maximum $\mathbf{z}$ axis value of the real 3D space in the standard coordinate system centered to the camera. Z is expressed in meters and quantized with 316 bins. We adopted S_2 for Panoptic and Human3.6m.

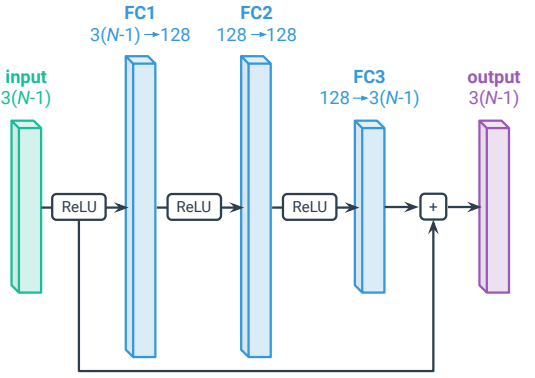

Figure 3.15: Structure of our Pose Refiner: 3 fully connected layers with ReLU activation followed by a skip connection. N is the number of joints. In all our experiments we considered $N = 14$.

Although the difference between these two spaces is minimal, we adopted S_1 for JTA because this dataset already provide a maximum camera distance, which is 100 meters.

Skeleton Grouping Details Let's consider K type of joints and $N_0, \dots, N_{K-1}$ number of detections for each joint type. Given N_0 different predicted heads, $\mathbf{j}_{0,0}, \dots, \mathbf{j}_{0,N_0-1} \in \mathbb{R}^3$, and $N_k, k \in [1, K-1]$ predicted joints of another type, $\mathbf{j}_{k,0}, \dots, \mathbf{j}_{k,N_k-1} \in \mathbb{R}^3$, we define $K-1$ cost matrices, $\mathfrak{D}_1, \dots, \mathfrak{D}_{k-1}$, as follows: $\mathfrak{D}_k : \{0, \dots, N_0 - 1\} \times \{0, \dots, N_k - 1\} \rightarrow \mathbb{R}$ where each element $d_{a,b}$ is defined as

$$d_{a,b} = \begin{cases} \|j_{0,a} - j_{k,b}\| & \text{if } \|j_{0,a} - j_{k,b}\| \leq 1.5 \cdot \tau_k \\ +\infty & \text{otherwise} \end{cases} \tag{3.12}$$

The threshold τ_k in (3.12) is the maximum distance between a head and a joint of type k (belonging to the same person) on the entire training set. For each $k = 1, \dots, K-1$, joint-head associations are made with the Hungarian algorithm using $\mathfrak{D}_k$ as cost matrix. The same joint-grouping procedure is applied on both multi-person datasets. By removing the anatomical constraints, results on Panoptic show an MPJPE degradation of about 9 millimeters while on JTA, no degradation in terms of metric has been observed.

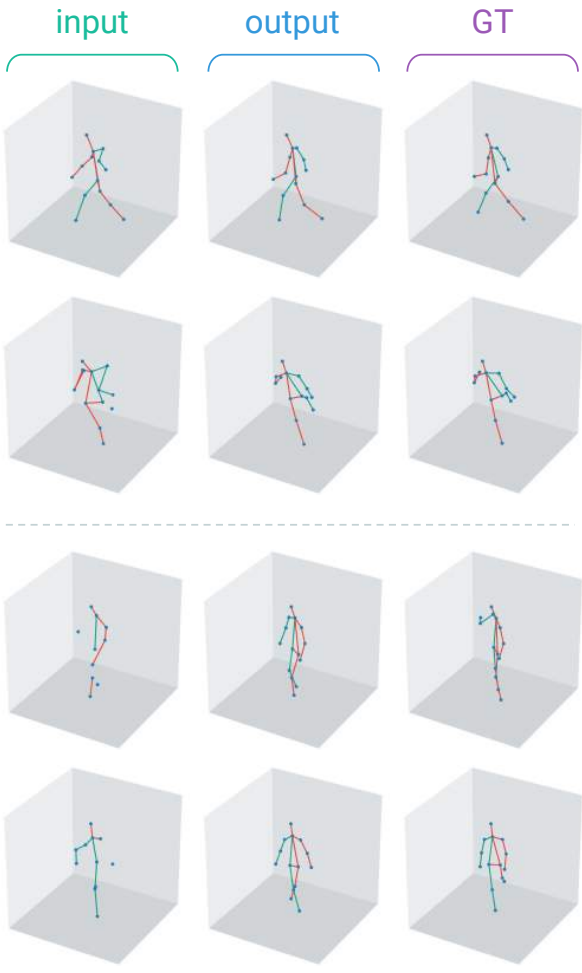

Figure 3.16: Qualitative results of our Pose Refiner model on the JTA dataset. 1^{st} and 2^{nd} rows: examples where the output is anatomically plausible and consistent with the ground truth; 3^{rd} and 4^{th} rows: examples where the output is anatomically plausible, but inconsistent with the ground truth.

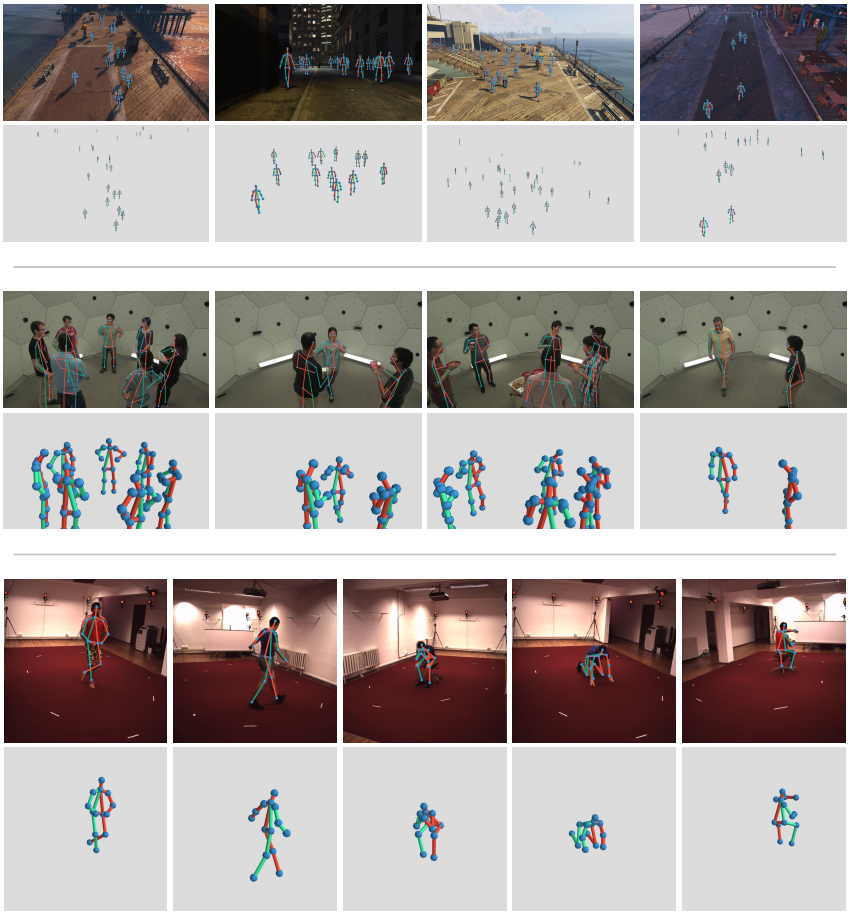

Figure 3.17: Additional qualitative results of our LoCO approach. 1st and 2nd rows: result of LoCO$^{(2)}$+ on the JTA dataset; 3rd and 4th rows: result of LoCO$^{(2)}$+ on the CMU Panoptic dataset; 5th and 6th rows: result of LoCO$^{(3)}$+ on the Human3.6m dataset

3.3 3D People Detection

The COVID-19 emergency has changed the way we live interpersonal social relationships, at work, in public and private spaces, in places of education, culture and leisure. The risk of contagion seems full-blown; until now, there are no conclusive studies which correlate environmental and endogenous factors with the greatest spread of the virus: instead, everything seems to correlate the contagion to proximity or to the contact between infected people and people susceptible to infection [7]. The spread of the infection seems to follow the epidemiological models that derive from the SIR models [38].

The phases that all the world is going to undertake after the lock-down will be characterized by living with the risk of contagion: the prerogative will be to take conscious and possibly interactive measures to minimise the possibility of contagion, while seeking a necessary resumption of social and working life.

Certainly, the IT technologies and in particular Artificial Intelligence can be valuable tools to monitor and predict risk levels in potentially crowded places.

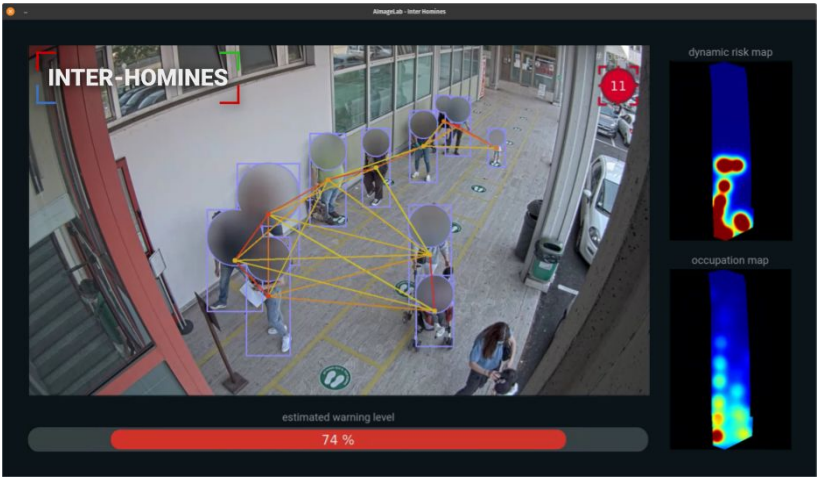

Figure 3.18: GUI of our system. In the main frame, anonymized bounding boxes are superimposed to the image. Colored links encodes people reciprocal distance. On the right, two maps shows the bird-eye view of the area. The estimated risk level of the scene resides at the bottom of the interface.

In fact, we propose an innovative and effective technological contribution based on Computer Vision and Deep Learning, in order to dynamic monitoring the acceptance of social distancing prevention measures through real-time calculation of the risk level, with particular reference to workplaces, public places and social areas. For statistical purposes, people behavior dynamics are stored in a database in a completely privacy compliant manner. The data can be used to identify the most critical areas and hours of the day in terms of number of people and risk level, in order to better address distance-related interpersonal prevention measures.

The system is called Inter-Homines (from the “Homo inter homines sum, capite aperto ambulo” - “I am a human among humans and I can walk with my face uncovered”) because people should be free to move and interact with uncovered faces while being safe at the same time.

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 "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.

Detecting people, their position in the space, their mutual distance is a typical application of Computer Vision. Many tools are available, using state-of-the-art deep learning architecture and geometry-based 3D reconstruction. Results are promising although still far to be applied everywhere by everyone. In this project, we can take advantage of a long term experience in computer vision for surveillance and people behavior understanding [63, 62], providing a novel detection pipeline running in real-time. It exploits standard fast camera calibrations, a people detector and a pose estimation methods.

Inter-Homines defines a model, validated by epidemiologists and parameterizable according to current regulations, which allows, in real-time, to associate each monitored area with: a) a space-time risk index, b) a dynamic safety level of the area, c) a dynamic map of interpersonal distances and d) a real time visualization of detected persons and distances. See Fig. 3.18 for the system output overview.

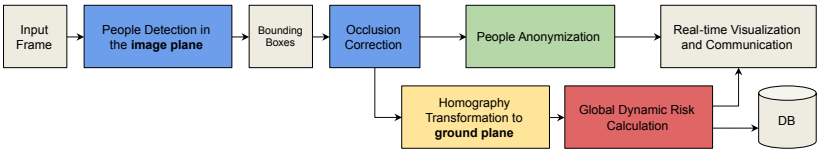

Figure 3.19: Schematization of the Inter-Homines pipeline: the input frame is processed to produce bounding box detections. Each detection is then refined by the Occlusion Correction module that copes with truncated bounding boxes. The image plane detection coordinates are then transformed to ground plane coordinates using an Homography Transformation. Those coordinates are then used to calculate the global risk. People coordinates and risk level are then stored into a database. Finally, the system outputs the anonymized frame along with the risk level and the risk maps.

3.3.1 Risk Model

After the outbreak of the COVID-19 pandemic, all the world learned the importance of the basic reproduction number, $\mathcal{R}_0$, as the statistical index indicating the degree of spread of the infection. In commonly used infection models, when $\mathcal{R}_0 > 1$ (in Italy has reached 4.3 during the spring of 2020) the infection will be able to start spreading in a population, but not if $\mathcal{R}_0 < 1$. Generally, the larger the value of $\mathcal{R}_0$, the harder it is to control the epidemic.

$\mathcal{R}_0$ is defined as the expected number of secondary cases produced by an infection in a completely susceptible population:

$$\mathcal{R}_0 = \alpha \cdot c \cdot d \tag{3.13}$$

where α is the transmissibility, c is the average rate of contact between susceptible and infected individuals, and d is the duration of infectiousness.

To understand if this quantity defines the epidemic threshold of a particular infection, we need to formulate a Susceptible-Infected-Removed (SIR) epidemic model [121]. This model deploys several assumptions: 1) closed population, 2) constant rates, 3) no births and deaths and 4) well mixed population.

Given a population of N individuals, let's consider S the number of susceptible people, I the infected, and R the removed. Removed people are those that cannot be infected, as they might have developed antibodies. Now let's define $s = \frac{S}{N}$,

$i = \frac{I}{N}$, $r = \frac{R}{N}$ as the fraction in each set. The SIR model is defied as:

$$\frac{ds}{dt} = -\beta si, \qquad \frac{di}{dt} = \beta si - vi, \qquad \frac{dr}{dt} = vi \tag{3.14}$$

An epidemic occurs if the number of infected increases: $\frac{di}{dt} > 0$. By considering that everyone is susceptible, we can substitute $s = 1$ obtaining the following inequality:

$$\alpha cd = \mathcal{R}_0 > 0 \tag{3.15}$$

$\mathcal{R}_0$ is essentially the entire theoretical basis of public health interventions for infectious diseases and it is simply the product of the transmissibility, the mean contact rate, and the duration of infection. In order to reduce transmissibility α we can develop vaccines, get people to use barrier contraceptives or use anti-retrovirals. To decrease mean contact c, the world decided to use isolation/quarantine, and health education programs. Finally, to reduce the duration of infection d, therapeutics and antibiotic treatment of bacterial infections that boost innate immune response can be exploited.

$\mathcal{R}_0$ is in generally computed as a posterior measure, but cannot be dynamically predicted in a robust way since the factor influencing $\mathcal{R}_0$ are not a priori easily measurable. In this work we cannot do anything a part from monitoring the acceptance of health education programs. In the past months, many countries decided the mandatory measures of security that concern the use of DPI and the social distance guidelines. Thus, in order to make c as small as possible, we should keep all the people at a distance larger that a threshold distance of a possible infection.

A viable way is to force people to stay in queues, maybe with some marker placed on the floor and with the constant attention of a human guard that controls the compliance of the social distancing norms. This is not always possible, especially in big malls and wide areas. Moreover, the human monitoring is not always optimal as the guard is subject to tiredness and lost of focus.

This is the reason why computer based systems joined with risk models can substitute human controllers and help to perform real-time monitoring of areas by assessing a level of possible risk, and, if necessary, giving a real time feedback to improve the safety and decrease the risk. In the following section, we propose a very simple model, that, using some thresholds validated by epidemiologists, models the dynamic risk in a given area.

The SIR model formulation, as described in the previous section, has validity when considering a population. Now, let's consider a much more restricted zone.

This could be an indoor area, like a waiting room of a public office, an entrance in a cinema or a shop. More precisely, let's consider a scene with N people $k_0, \ldots, k_{N-1}$ at a given time t. Given two subjects k_i and k_j with distance $d_{i,j}$, we define their reciprocal risk as follows:

$$rr_{i,j}^{(t)} = \eta\, e^{-\beta \max(0,\, d_{i,j} - \tau)} \tag{3.16}$$

where η, β and τ are parameters that respectively control height, slope and the full width at maximum of the function. In this specific application, η is a mitigator used to decrease the risk when some criteria are met, e.g., when at least one of the two people is wearing a facial mask. β, instead, controls how the risk decreases when the distance is greater than τ and can model environmental characteristics such as air temperature and the presence of air conditioning. Lastly, τ, controls the transmissibility of the disease via respiratory droplets and define the minimal distance allowed between two people. It should follow World Health Organization and national guidelines but can be further increased to better preserve the safety of people in critical places such as COVID-19 hospital units. We then define the individual risk at time t as:

$$R_i^{(t)} = \max_{j=0\ldots N-1, j\neq i} \{rr_{i,j}^{(t)}\} \tag{3.17}$$

The global risk at t of the scene is then computed as follows:

$$G^{(t)} = \min\left(1,\ \frac{1}{C} \sum_{i=0}^{N-1} R_i^{(t)}\right) \tag{3.18}$$

where C is the maximum capacity of the scene. This capacity can be either given by the user or calculated using simple covering algorithms. Finally. the dynamic global risk is computed as:

$$D^{(t)} = \frac{1}{W} \sum_{w=0}^{W-1} G^{(t-w)} \tag{3.19}$$

where W is the size of the temporal window. At a given time t, $D^{(t)} \in [0, 1]$ is the global risk of the scene and it is used to trigger alarms when it reaches a given threshold.

3.3.2 Method

Here we give an overview of the pipeline we used to process videos in real time. The aim of our Inter-Homines system is to detect people, compute their distance and provide a dynamic risk 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. See Fig. 3.19 for a schematization of the pipeline.

The following subsections summarize the key elements of our system. Section 3.3.2 describes the people detection stage and elaborates on its challenges. Section 3.3.2 illustrates our proposed keypoint localization solution which addresses the occulsion problem peculiar of surveillance scenarios and also provide the head position for anonymization purposes. Next, in Section 3.3.2, we describe how we convert points from image plane to ground plane and, finally, Section 3.3.2 illustrates the system outputs.

People Detection

As we are interested in the best speed-accuracy trade-of, we choose CenterNet [304] as a people detector. In particular, we rely on the DLA backbone [288] which yields 51.3% AP for the people class on MS COCO [149], running at 52 FPS on a Titian XP.

Let $I \in R^{W \times H \times 3}$ be the input image having width W and height H. CenterNet outputs a keypoint heatmap $\hat{Y} \in [0,1]^{\frac{W}{R} \times \frac{H}{R} \times C}$, where $R = 4$ is the output stride and C is the number of keypoint types. Keypoint types include $C = 80$ object categories but in this work we only consider the "people" class. Detected keypoints corresponds to a prediction $\hat{Y}_{x,y,c}, = 1$ and 0 otherwise. To recover the discretization error generated by the output stride, CenterNet further predicts a local offset $\hat{O} \in \mathcal{R}^{\frac{W}{R} \times \frac{H}{R} \times 2}$ for each center point.

Let $(x_1^{(k)}, y_1^{(k)}, x_2^{(k)}, y_2^{(k)})$ be the bounding box of object k of the "people" class and $p_k = (\frac{x_1^{(k)}+x_2^{(k)}}{2}, \frac{y_1^{(k)}+y_2^{(k)}}{2})$ it's center point. CenterNet predicts all center points for each object k and further regresses to the object size $s_k = (x_2^{(k)} - x_1^{(k)}, y_2^{(k)} - y_1^{(k)})$.

At running time, we first extract the peaks in the heatmap for the "people" category. We consider all the responses whose value is greater or equal to its 8-connected neighbors. Let $\hat{\mathcal{P}} = \{(\hat{x}_i, \hat{y}_i)\}_{i=1}^{n}$ be the set of n detected center points where keypoint values $\hat{Y}_{x_i y_i c}$ are utilized as a measure of its detection

Figure 3.20: Examples of CenterNet bounding boxes (pink), refined bounding boxes and head localization (green).

confidence. Bounding boxes are produced at location:

$$\begin{aligned}(\hat{x}_i + \delta\hat{x}_i - \hat{w}_i/2,\ \ \hat{y}_i + \delta\hat{y}_i - \hat{h}_i/2,\\ \hat{x}_i + \delta\hat{x}_i + \hat{w}_i/2,\ \ \hat{y}_i + \delta\hat{y}_i + \hat{h}_i/2),\end{aligned} \tag{3.20}$$

where $(\delta\hat{x}_i, \delta\hat{y}_i) = \hat{O}_{\hat{x}_i,\hat{y}_i}$ is the predicted offset and $(\hat{w}_i, \hat{h}_i) = \hat{S}_{\hat{x}_i,\hat{y}_i}$ is the predicted size. Since the prediction are directly produced from the keypoint estimation, there is no need for IoU-based NMS or other post-processing techniques. This makes CenterNet faster w.r.t. other detectors, making it suitable for real time applications.

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. As shown in Fig. 3.20 (pink bounding boxes), 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 homograpy, we need to know the exact position (in image plane) of the feet of each detected person. This task cannot be accomplished by solely relying on CenterNet.

Feet and Head Localization

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 but effective CNN that, given an image M tightly containing a person, it regresses to the midpoint $P_f = (x_f, y_f)$ of the segment having the two feet as endpoints. This ensures that we know the exact position in image plane where every person touches the ground. Since we are also interested in anonymizing the face of each detected person, we further predict the location of the head $P_h = (x_h, y_h)$.

We replaced the last 1000 class classification layer of Resnet50 [94] with two heads composed by an adaptive average pooling layer and a fully connected layer with output size equal to 2. The adaptive average pooling takes care of the difference in size that each bounding box fed to the network can have. Training has been carried out for 10 epochs using an MSE loss with Adam optimizer, batch size of 64 and learning rate 0.001.

We used JTA [63] as the training dataset since it is the only surveillance dataset available in 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 behaviours. This step ensures a more precise localization of the feet while also coping with truncated bounding boxes. As shown in Fig. 3.20 (green 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.

From Image Plane to Ground Plane

The camera projection matrix P is a 3×4 matrix which describes the mapping of a pinhole camera [92] from 3D points in the world to 2D points in an image. Let X be a representation of a 3D point in homogeneous coordinates (a 4-dimensional vector), and let y be a representation of the image of this point in the pinhole camera (a 3-dimensional vector), we have $y = PX$. The camera projection matrix can be decomposed as:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x & 0 \\ 0 & f_y & c_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} r_{1,1} & r_{1,2} & r_{1,3} & t_X \\ r_{2,1} & r_{2,2} & r_{2,3} & t_Y \\ r_{3,1} & r_{3,2} & r_{3,3} & t_Z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{3.21}$$

where the intrinsic parameters f_x, f_y and c_x, c_y are the camera focal length and principal points respectively while $r_{i,j}$ and t_i are the extrinsic parameters which define the rotation and the translation used to describe the rigid motion of an object in front of a still camera. Finally, u and v are the coordinates of the projected point in pixels while X, Y and Z are the coordinates of a 3D point in the world coordinate space. By considering the simpler case of a projection of planar points, where each 3D point lies on the same plane (e.g. the ground), we can simplify the formulation considering $Z = 0$. For planar surfaces, 3D to 2D perspective projection reduces to a 2D to 2D transformation:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{1,1} & r_{1,2} & t_1 \\ r_{2,1} & r_{2,2} & t_2 \\ r_{3,1} & r_{3,2} & t_3 \end{bmatrix} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} \tag{3.22}$$

and by doing the products we finally obtain the planar homography matrix H. The planar homography relates the transformation between two planes (e.g. the image plane and the ground plane):

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} h_{1,1} & h_{1,2} & h_{1,3} \\ h_{2,1} & h_{2,2} & h_{2,3} \\ h_{3,1} & h_{3,2} & h_{3,3} \end{bmatrix} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} = H \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} \tag{3.23}$$

Since H maps from ground plane to image plane, but we are interested in the opposite transformation (from image plane to ground plane), we now need to calculate the inverse homography matrix H^{-1}. An homography matrix H is always invertible, and its inverse is still an homography transformation:

$$\begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} = H^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \tag{3.24}$$

A practical way to calculate the homography matrix H of Eq. 3.23 is to find a set of at least four points pairs of target and source planes and to minimize the

Table 3.13: 3D detection results on JTA Dataset. In PR (precision), RE (recall) and F1, @t indicates that a predicted person is considered "true positive" if the distance from the corresponding ground truth location is less than t. The max range indicates the maximum distance considered in the calculation.

		PR	RE	F1	PR	RE	F1	PR	RE	F1
max range		@0.5 m			@1.0 m			@1.5 m		
10m	w/o Occ. Corr.	83.3	78.2	80.0	90.4	85.2	87.0	92.5	87.6	89.3
	Full Pipeline	88.0	84.7	85.5	92.9	89.2	90.2	94.2	90.5	91.5
20m	w/o Occ. Corr.	69.1	59.7	63.4	85.3	73.8	78.3	91.8	79.7	84.4
	Full Pipeline	74.9	66.9	70.0	88.8	79.2	82.8	93.6	83.6	87.4
30m	w/o Occ. Corr.	59.4	46.8	51.5	77.3	60.7	66.9	85.8	67.2	74.1
	Full Pipeline	65.3	53.0	57.5	81.2	65.4	71.1	88.3	70.9	77.3
100m	w/o Occ. Corr.	53.7	31.6	38.0	71.3	41.7	50.4	80.3	46.8	56.7
	Full Pipeline	60.9	36.2	43.3	77.1	45.2	54.5	84.7	49.4	59.7

back-projection error:

$$\sum_{i=0}^{N}\left[\left(u_i-\frac{h_{1,1}X_i+h_{1,2}Y_i+h_{1,3}}{h_{3,1}X_i+h_{3,2}Y_i+h_{3,3}}\right)^2+\left(v_i-\frac{h_{2,1}X_i+h_{2,2}Y_i+h_{2,3}}{h_{3,1}X_i+h_{3,2}Y_i+h_{3,3}}\right)^2\right] \tag{3.25}$$

However, if not all of the point pairs fit the rigid perspective transformation, this initial estimate will be poor. To solve this problem we employ the RANSAC iterative method, trying many different random subsets of the corresponding point pairs (of four pairs each). We then estimate the homography matrix applying a simple least-square algorithm using this subset, and then compute the quality of the computed homography, which is the number of inliers. The best subset is then used to produce the initial estimate of the homography matrix. The computed homography matrix is refined further (using only the inliers) with the Levenberg-Marquardt method to further reduce the re-projection error. The homography matrix is determined up to a scale. Thus, it is normalized so that $h_{3,3} = 1$.

This method of using an homography transformation to obtain 3D coordinates is the most appropriate when we want to monitor an approximately flat area (such as a town square) using a fixed camera and there is the possibility of making appropriate measurements in the monitored space.

To easily obtain the points pairs of image and ground planes needed to find

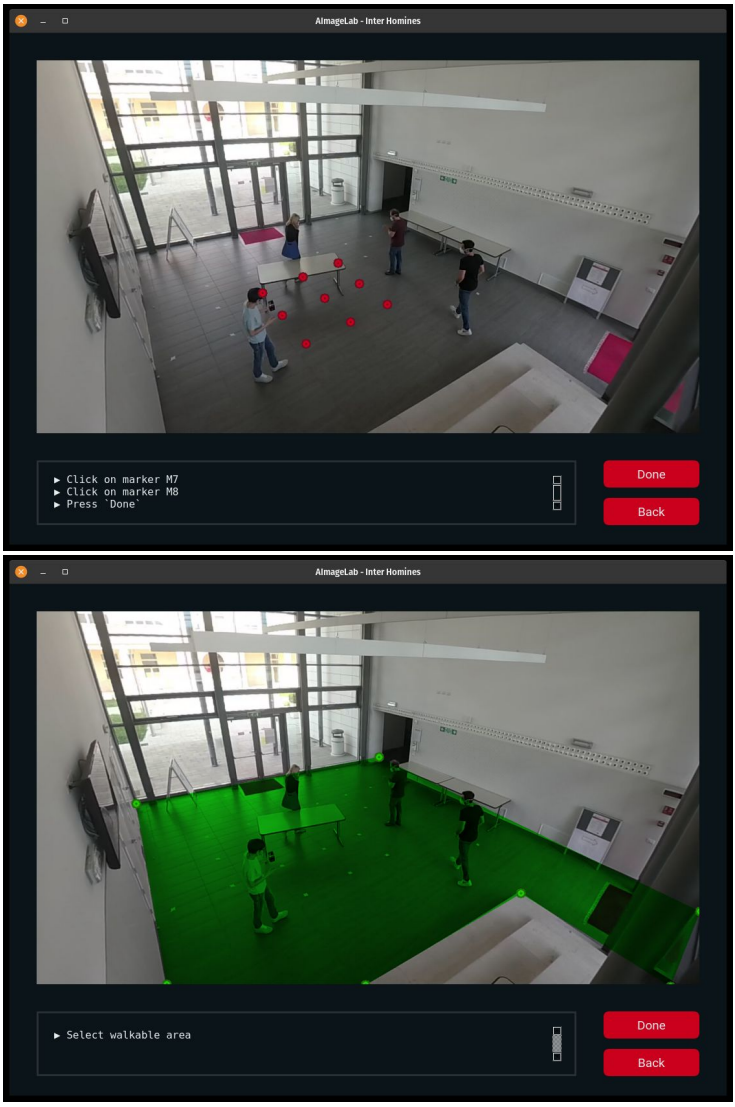

Figure 3.21: GUI used during system calibration for homography matrix calculation (top) and for walking area selection (bottom).

the homography matrix H, we designed a simple procedure that we call "system calibration". This procedure consists in placing nine markers at the center of the monitored area, fully visible from the camera. The markers are placed in a grid pattern as in Fig. 3.21. By means of a simple graphic interface, the user can take a snapshot of the camera and click with the cursor the centers of the nine markers in order to acquire the pixel coordinates. In practice, we utilize a special carpet with the nine markers printed on it. The use of the carpet automates the real world measurements as we already know the distance between markers in the carpet, making the system calibration fast, less prone to errors and feasible by everyone. Once the nine pairs of points have been identified and the homography matrix calculated, the carpet can be safely removed and the system will continue to work properly as long as the camera maintains its position.

During the system calibration an optional procedure of selecting the "walking area" can be carried out. Again, a simple graphic interface let the user draw a polygon on the snapshot taken from the camera, as shown in Fig. 3.21. The pixel vertices are then converted to ground coordinates that are used to exclude detections whose 3D position is outside the walking area. This is useful, for example, to ignore mirrors or windows that can reflect people causing unwanted detections.

Given a bounding box of a person, we can now extract its central point (u, v) of the lower side of the box (i.e. the image coordinate where the person touches the ground with his feet), and utilize Eq 3.24 to obtain the corresponding (X, Y) point in ground plane. Now that we have the 3D position of every person in the scene, the dynamic global risk in Eq. 3.19 can be calculated and given as output along with other information that we summarize in the following subsection.

System Output

A convenient graphical interface highlights all the main results of the analysis of our Inter-Homines system, allowing to evaluate at a glance the crowding conditions in the monitored area (see Fig. 3.18). This interface is made with Qt to guarantee compatibility with all the main operating systems.

Anonymized Frame It shows real-time the bounding box detections superimposed to the input RGB frame. The system is privacy compliant and all the faces are obscured. Colored segments connect people who are at an estimated distance lower than a defined upper threshold distance (typically 3 m). The color indicates

the extent of the infraction, going from a dark red for the most serious infraction to yellow for the minor ones.

People Counter At the top right of the frame we also display the number of detected people updated in real time. This number is an average computed in a window of W frames to account for miss detections and false positives.

Dynamic Risk and Occupation Maps In the right part of the interface two bird eye views of the walking area are updated real-time. The Dynamic Risk map shows a snapshot of the current situation of the area. The Occupation map, instead, displays the overall aggregated risk and it is computed by averaging the Dynamic Risk maps of the whole day. It is used to identify areas with a larger risk for statistical and predictive purposes. Note that people outside of the walking area are completely ignored and do not affect the statistics.

Estimated Warning Level In the lower part of the window a bar represents the total estimated risk in the monitored area and it is computed using Eq. 3.19. The application provides the possibility to send an alarm signal (example: send an email / audio notification) if certain thresholds on the number of people or on the risk level are exceeded. The thresholds and the notification methods of their exceeding will be defined according to the needs of the context in which the system will operate.

Weekly Report Since we want to give insightful statistics to help with the prevention of the infection, our system periodically produce a report. The report contains statistics about number of people, risk level, number of infractions and occupation maps aggregated by hours and days. To this end we utilize a non-relational database to store timestamp and position of each person captured by our system.

3.3.3 Experiments

In order to validate the effectiveness of our system, we performed a series of experiments leveraging JTA [63]. JTA is a massive dataset for pedestrian pose estimation and tracking in urban scenarios created exploiting the highly photorealistic video game *Grand Theft Auto V*. The videos feature a vast number of different people appearances, in several urban scenarios at varying illumination conditions

Figure 3.22: Examples from the JTA dataset exhibiting its variety in viewpoints, number of people and scenarios. Ground truth bounding boxes are superimposed to the original images. Green color is used for people having a distance from the camera between 0 and 20 meters, yellow for people between 20 and 40 meters and red for people between 40 and 100 meters.

and viewpoints. Each clip comes with a precise annotation of visible and occluded body parts, people tracking with 2D coordinates in image plane and 3D coordinates in camera space. JTA overcomes all the limitation of existing datasets in terms of number of entities and available annotations. Each video contains a number of people ranging between 0 and 60 with an average of more than 21 people, totaling almost 10M annotated body poses over 460,800 densely annotated frames. The distance from the camera ranges between 0.1 and 100 meters, resulting in pedestrian heights between 20 and 1100 pixels. JTA is composed by a set of 512 Full HD videos, 30 seconds long, recorded at 30 fps.

As shown in Fig. 3.22, despite being a synthetic dataset, JTA features highly challenging and complex situations, peculiar of surveillance scenarios, where people are often dominated by severe body part occlusions and truncations. We believe this dataset is the perfect choice to validate a system that targets global safety.

Since we can not perform the system calibration procedure on an already recorded dataset, i.e. we can not physically place the markers at the center of

the scene, we designed a simple heuristic to directly recover the nine points pairs using the dataset annotations. With the assumption that every foot of each person lies on the same plane, for each JTA sequence, we linearly regressed the ground plane utilizing the 3D coordinates of every foot in every frame of that sequence. By recovering a unit normal vector of the plane and two orthonormal vectors lying on the plane we were able to find the orthonormal base of the new space that allowed us to move each 3D coordinate into a space where each foot has the same y coordinate (according to the standard camera system). Now, since each foot coordinate has the same y, we can get rid of it and considering the new (x, z) coordinates as ground coordinates. As we are interested in nine points pairs of target and source planes, we utilized a K-Means implementation to find nine foot cluster centers. Utilizing a clustering method ensures that the nine points are far from each other. Once recovered the foot cluster centers, we remapped those coordinates into the original standard camera space and projected them into the image plane using the pinhole camera model. The 2D projected coordinates and the 2D foot clusters now form the nine points pairs needed to calculate the homography matrix.

Experiments are conducted on every 10^{th} frame of a subset of the JTA test set where we carefully removed the sequences that contain camera motion and people at different heights, e.g. people going up the stairs, as our method assumes static camera and flat ground plane. Tab. 3.13 shows the precision, recall and F1 obtained using different thresholds and considering different camera distance ranges. As the range increase, we observe a decrease in performances, due to the fact that small people are hardly detected and homography transformation becomes less reliable. Since we are interested in evaluating the impact of that occlusions have in the performance of our system, we reported the results with and without the occlusion correction module. As can be shown, the correction is always beneficial, especially when people are close to the camera.

To better understand how performance degrades as distance increases, in Fig. 3.23, first row, we plotted the F1 score at different thresholds w.r.t. the camera distance. It is interesting to note that performance worsens when people are too close to the camera. In Fig. 3.23, second row, we plotted the same quantity but, this time, the F1 score is calculated considering all the people with distance less than the camera distance, and not equal to the camera distance.

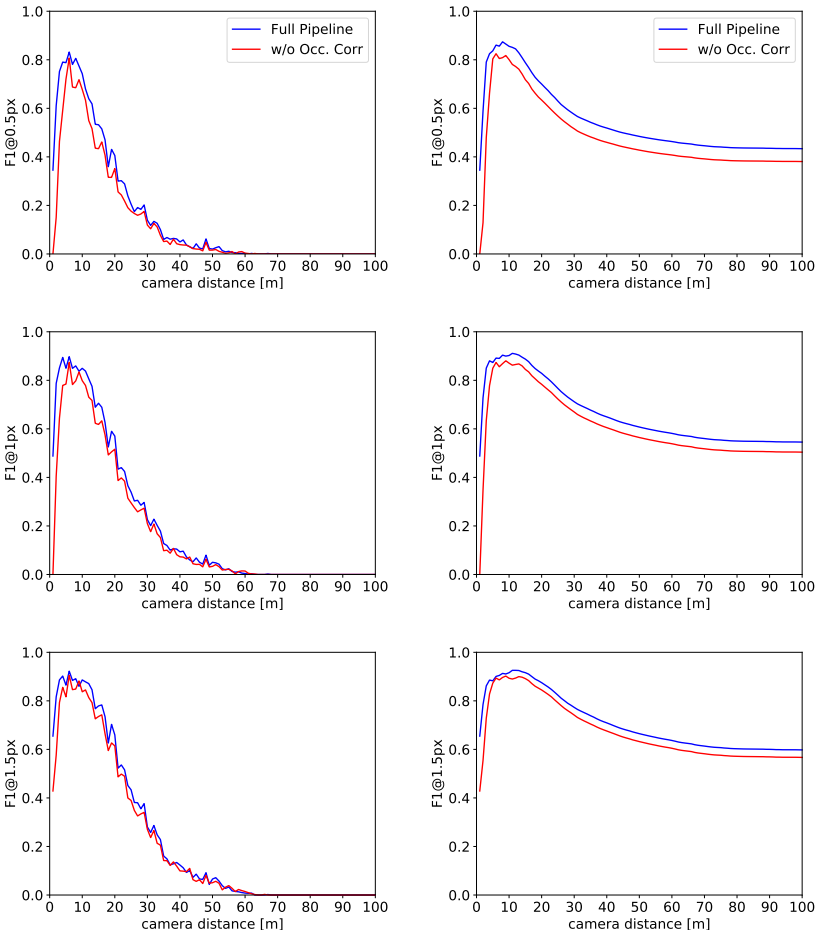

Figure 3.23: F1 score vs. camera distance at different thresholds (first column) and F1 score vs. max camera distance at different thresholds (second column).

3.4 Attribute Recognition

While recent efforts in people detection, recognition, and tracking enabled a plethora of video-surveillance applications, e.g. people identification by [159], pose estimation by [224] and action analysis by [97], occlusion is still an open problem. The occlusion issue is well known in the people detection and tracking literature and generally affects any intelligent video surveillance system, but it is debatable whether a real solution to the problem could exist effectively. In fact, whenever an occlusion occurs we observe a removal of information from the observed scene. The occluded portion of an object, indeed, becomes unknown and, in a Parmenidean sense, it does not exist until it can be observed. For this motivation, most of the literature focused on counteracting the phenomenon conveying occlusion robustness to either detection, tracking, or re-id systems as by [184, 145, 196, 269, 41]. In the matter of fact, recovering the image content from an occlusion is feasible only in the case where the target has been previously observed e.g. in a video stream. This is the approach followed also by many tracking solutions which memorize several detected appearance of the person, to discard occlusions as "less frequent accidents" w.r.t. the normal visible appearance. Nevertheless, leveraging on the generative capabilities of GANs by [83], we aim at demonstrating that it is indeed possible to hallucinate a plausible representation of the occluded content even when it has never been previously observed, i.e. in single images.

Following on our previous work on the topic ([60]), in this work, we introduce a novel network that leverages the generative power of GANs for hallucinating the occluded portion of the image without any guidance of an attention mechanism that could provide instance level information about the occluding person. The proposed solution aims at generating or reconstructing the image of a person which could be plausible in many senses: a) similar to images of real people, observed in the training dataset; b) acceptable at pixel level as a person shape; c) capable to preserve similar visual attributes of the ground truth de-occluded image. This is carried out by exploiting solutions for attribute classifications (e.g. male/female, young/old, with/without trousers, etc.) and integrating them in a U-net like generative and adversarial architecture.

Another major problem that arises when dealing with occlusions, through learning-based solutions, is the lack of large-scale datasets providing realisticly occluded and non-occluded pairs of images. Most of the proposed solution in literature, like the ones introduced by [60, 194, 192] paste together different people detections, or manually add random objects or textures to a non-occluded image.

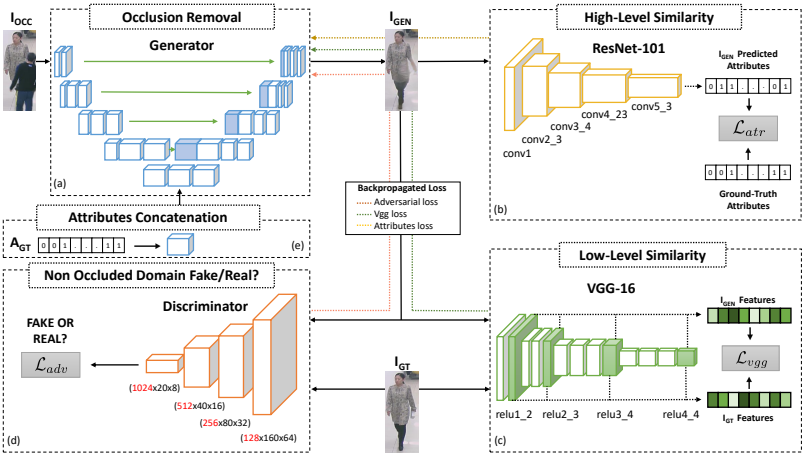

Figure 3.24: A schematic representation of the training procedure adopted in our work. The Generator (a) takes the occluded image I_{OCC} as input and the attributes of the person A_{GT} (e) as a further conditioning element. The output of the Generator I_{GEN} is the restored image, with no occlusion. To train the Generator, we fed I_{GEN} to three different networks: ResNet-101, VGG-16, and the Discriminator. (b) ResNet-101 gives a prediction of the I_{GEN} attributes which are compared with the ground-truth ones for the $\mathcal{L}_{atr}$ computation, in order to maximize high-level similarity. (c) The feature maps extracted from different layers of VGG-16 are used to calculate the content loss between I_{GEN} and I_{GT} with the aim of encouraging low-level similarity. (d) 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. The Discriminator is trained alongside the Generator to distinguish between generated "fake" images and "real" ones. At evaluation time, only the Generator network and the Attributes Concatenation are used.

These processes ultimately fail to generate realistic data and are thus a liability when employed for training a neural network that aims at resolving the occlusion while keeping the rest of the image coherent (e.g. the background) and preserving the person's attributes. To address this issue, we propose a novel, fully automatic, way to generate realistic occlusion pairs by exploiting the recent achievements in object segmentation by [93]. These results are high-fidelity occlusion pairs, where the background of the original image is preserved and the generated occlusion is more realistic. Additionally, thanks to the software provided by [63], we created a massive computer graphics generated dataset[1], in which we artificially created a large collection of occluded pedestrians. Additionally, we recovered from the game engine their attributes by manually annotating just the models. To our knowledge, this is the first CG dataset for the purpose of de-occluding people having a set of annotated person attributes (e.g. sex, hair color, dress style, etc.).

To summarize, our contributions are threefold:

- We propose a novel generative adversarial network that is able to solve occlusions in pedestrian images by hallucinating the missing parts while keeping both the appearance and the background coherent;
- We devise a new way for synthetically generating occlusion pairs that result in more realistic images when compared to other methods previously employed, also by creating a huge CG dataset;
- We propose a method for conditioning the occluded body part restoration on pedestrian attributes and consequently improving the generation process.

We show by experiments that the design of a conditional GAN that is aware of the attributes can acceptably hallucinate pedestrian and we experimentally demonstrate that this information is helpful in guiding the generation process. Results are interesting in terms of very high accuracy, outperforming other previous methods. We believe that our method could be useful in many computer vision systems, from surveillance, automotive to human behavior understanding tasks.

3.4.1 Method

The goal of our work is to reconstruct occluded body parts of pedestrians in different surveillance scenarios. Given an image of an occluded pedestrian as the network input, we aim at removing the obstructions and replacing them with body

[1] Leveraging on the highly photo-realistic graphics of GTA V video-game.

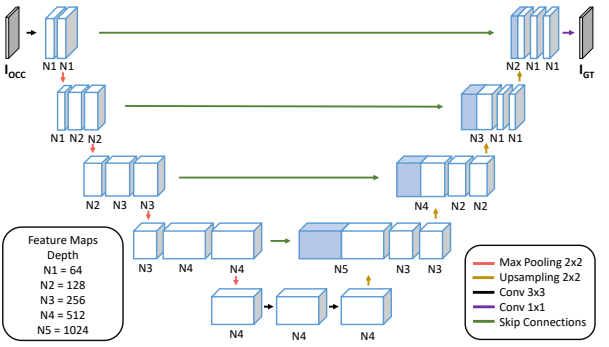

Figure 3.25: Architecture of our Generator network with corresponding number of feature maps and kernel sizes. The figure also depicts max-pooling and upsampling operations, along with skip connection gates.

parts that could likely belong to the occluded person. Note that, differently from the task of inpainting, we don't want to guide the network with the information about what portion of the image we want to remove and complete. For this purpose, we want to learn an image transformation between pairs of occluded images I_{OCC} and not occluded images I_{GT}.

In order to accomplish this, we propose the training procedure depicted in Fig. 3.24: our pipeline takes as input the occluded image I_{OCC}, along with the relative attributes A_{GT} and outputs the restored image I_{GEN}. I_{GEN} is then inputted to the three networks ResNet-101, VGG-16, and the Discriminator in order to compute the three components of our loss. Each loss component is then backpropagated through the input, updating only the Generator's weights.

More precisely, to achieve a full body restoration, we train the Generator network G as a feed-forward CNN G_{θ_g} with parameters θ_g. For N training pairs images (I_{OCC}, I_{GT}) and their relative attributes A_{GT}, we solve:

$$\hat{\theta}_g = \arg\min_{\theta_g} \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}_{total} \left(G_{\theta_g} \left(I^n_{OCC}, A^n_{GT} \right), I^n_{GT} \right) \tag{3.26}$$

Here $\hat{\theta}_g$ is obtained by minimizing the loss function $\mathcal{L}_{total}$ described in the next subsection. As a result, our generator network learns a mapping from observed images I_{OCC} to output image I_{GEN}. This differs from what did by [108] and [177] which use random noise along with the input image.

Table 3.14: Classification performances of our ResNet-101 on RAP dataset

Method	mA	Accuracy	Precision	Recall	F1
ACN [249]	69.66	62.61	**80.12**	72.26	75.98
DeepMAR [140]	73.79	62.02	74.92	76.21	75.56
DeepMAR* [141]	74.44	63.67	76.53	77.47	77.00
HP-Net [153]	76.12	65.39	77.33	78.79	78.05
ACN-Res50 [60]	**79.73**	64.13	76.96	78.72	77.83
Ours	78,66	**66,23**	77.85	**79.71**	**78.77**

Table 3.15: Detailed comparison between various pedestrian attribute datasets

Dataset	# Scenes	# Samples	# Attributes	Min Res.	Max Res.
PETA [50]	-	19,000	61(+4)	17×39	169×365
RAP [141]	26	41,585	69(+3)	36×92	344×554
PA-100K [153]	598	100,000	26	50×100	758×454
AiC	512	125,000	24	35×85	602×1080

Following [83], we further define the Discriminator network D_{θ_d} with parameters θ_d, that we train alongside G_{θ_g} with the aim of solving the adversarial min-max problem:

$$\min_G \max_D \mathbb{E}_{I_{GT} \sim p_{data}(I_{GT})}[\log D\left(I_{GT}\right)] + \mathbb{E}_{I_{OCC} \sim p_{gen}(I_{OCC},)}[\log 1 - D\left(G\left(I_{OCC}, A_{GT}\right)\right)] \quad (3.27)$$

where $D(I_{GT})$ is the probability of I_{GT} being a "real" image while the component $(1 - D(G(I_{OCC}, A_{GT}))$ is the probability of $G(I_{OCC}, A_{GT}) = I_{GEN}$ being a "fake" image. The min-max formulation force G to fool the D, which is adversarially trained to distinguish between generated "fake" images and "real" ones. Thanks to this approach, we obtain a Generator network capable of learning solutions that are similar to not occluded images thus indistinguishable by the Discriminator. Note also that, differently from what did by [108], we do not concatenate input images I_{OCC} to the "fake" images I_{GEN} or to the "real" images I_{GT} as Discriminator input.

Loss Function

The definition of the loss function $\mathcal{L}_{total}$ is crucial for the effectiveness of our Generator network. We propose the following loss formulation, composed by a weighted combination of three components:

$$\mathcal{L}_{total} = \overbrace{\underbrace{\mathcal{L}_{adv}}_{\text{adver. loss}} + \underbrace{\lambda_1 \cdot \mathcal{L}_{vgg}}_{\text{cont. loss}} + \underbrace{\lambda_2 \cdot \mathcal{L}_{atr}}_{\text{attr. loss}}}^{\text{total loss}} \tag{3.28}$$

The intuition behind this loss formulation is that we want the generated images to contain real people (thanks to $\mathcal{L}_{adv}$), to have similar feature representations (thanks to $\mathcal{L}_{vgg}$) and to preserve visual attributes (thanks to $\mathcal{L}_{atr}$) w.r.t. their non occluded ground truth versions.

The first term of Eq. (3.28) is the adversarial loss $\mathcal{L}_{adv}$. This component encourages the Generator network G to generate images belonging to the not occluded domain of pedestrians by fooling the Discriminator network D. The adversarial component relative to the Generator network is calculated as follows:

$$\mathcal{L}_{adv} = -\sum_{n=1}^{N} \log D\left(G\left(I_{OCC}, A_{GT}\right)\right) \tag{3.29}$$

Where $D(G(I_{OCC}, A_{GT}))$ is the probability that $G(I_{OCC}, A_{GT})$ is classified as "real" by the discriminator network. Minimizing $\log D\left(G\left(I_{OCC}, A_{GT}\right)\right)$ instead of $\log D\left[1 - \left(G\left(I_{OCC}, A_{GT}\right)\right)\right]$ is preferred in order to reach a better gradient behavior as indicated by [83]. As a possible drawback, the images produced by the Generator network G are forced to be realistic thanks to the Discriminator network D, but they can be unrelated to the original input. For instance, the output could be a plausible image of a pedestrian displaying a very different aspect w.r.t. the input image. Thus, is essential to mix the adversarial loss $\mathcal{L}_{adv}$ with an additional loss, such as L1 or L2, that evaluate the per-pixel distance between the generated and the ground truth image. Usually, training a network using such losses leads to reasonable solutions. However, the outputs appear blurred with lack of high-frequency details.

A possible solution for generating sharper results is to adopt a different content loss, like the perceptual loss introduced by [113] and used also by [135] and [130]:

$$\mathcal{L}_{vgg(i,j)} = \frac{1}{W_{i,j} H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} \left(\phi_{i,j}(I_{GT})_{x,y} - \phi_{i,j}(I_{GEN})_{x,y}\right)^2 \tag{3.30}$$

Figure 3.26: Examples from the AiC dataset exhibiting its variety in viewpoints, illuminations and scenarios.

where $W_{i,j}$ and $H_{i,j}$ are the dimensions of the feature maps $\phi_{i,j}$ obtained by the j-th convolution after ReLU activation and before the i-th max-pooling layer within the VGG16 network, pre-trained on ImageNet by [49], as done by [113].

The $\mathcal{L}_{vgg}$ that we employed in our work is based on the sum of different intermediate layers of VGG16:

$$\mathcal{L}_{vgg} = \sum_{i,j \in L} \mathcal{L}_{vgg(i,j)} \tag{3.31}$$

where $\mathcal{L}_{vgg(i,j)}$ is taken from eq. 3.30 and L is the set of used activations. Rather than encouraging the pixels of the output image I_{GEN} to exactly match the pixels of the target image I_{GT}, we instead encourage them to have similar feature representations as computed by the VGG16 network. As demonstrated by [113] and [160], minimizing the content loss for higher layers do not preserve color and textures. As we reconstruct from early layers, instead, images tend to be perceptually similar to the target image I_{GT} in terms of color and texture.

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 $\mathcal{L}_{atr}$. Like for the perceptual loss $\mathcal{L}_{vgg}$, we used a classification network as loss function. In particular, we adapted ResNet-101 by [95], pre-trained on ImageNet, to the task of multi-attribute classification. More

precisely, we replaced the last two layers (the average pooling and the last fully connected layer) in order to fit the desired input shapes. 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. In this case, we used a weighted binary cross entropy:

$$\mathcal{L}_{atr} = -\sum_{i=1}^{N_A} \exp\left(1 - r_i\right) \cdot \left(y_i \cdot \log\left(\psi_i(I_{GEN})\right)\right) + \exp\left(r_i\right) \cdot \left(1 - y_i\right) \cdot \log\left(1 - \psi_i(I_{GEN})\right). \quad (3.32)$$

Here, N_A is the number of attributes classified by the ResNet-101, r_i is the positive ratio of i-th attribute. ψ is the output of our attribute classification network and y_i is the i-th ground truth label.

Networks Architecture

Generator Network Our Generator structure differs from those presented by [210] and [60]: following [227] and [108] we propose the "U-Net" like architecture depicted in Fig. 3.25. In particular, the structure of our network slightly differs from the one described by [227] and [108]. The network is composed by 4 down-sampling blocks and a specular number of up-sampling components. Each down-sampling block consists of 2 convolutional layers with a 3×3 kernel. Each convolutional layer is followed by a batch normalization and a ReLU activation. Finally, each block has a max-pooling layer with stride 2. The up-sampling part has a very similar but overturned structure, where each block is composed by an up-sampling layer of stride 2. After that, each block is equipped with 2 convolutional layers with a 3×3 kernel. The last block has an additional 1×1 kernel convolution which is employed to reach the desired number of channels: 3 RGB channels in our case. A $tanh$ has been used as final activation. We additionally inserted skip connections between mirrored layers, in the down-sampling and up-sampling streams, in order to shuttle low-level information between input and output directly across the network. Eventually, padding is added to avoid cropping the feature maps coming from the skip connections and concatenate them directly to the up-sampling blocks outputs. Roughly speaking, our task can be seen as a particular case of image-to-image translation, where a mapping is performed between the input image and the output image. Additionally, for the specific problem we are considering, input and output share the same underlying structure despite differing in superficial appearance. Therefore, a rough alignment is present between the

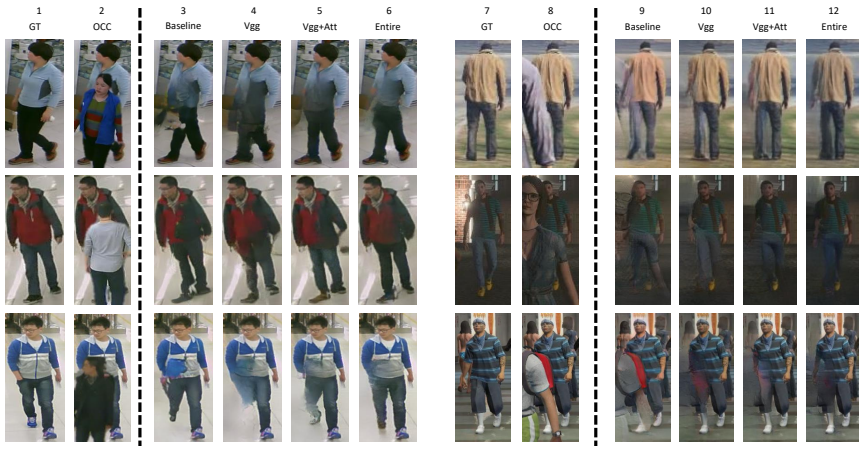

Figure 3.27: Qualitative results based on the ablation study on RAP dataset (leftmost) and AiC dataset (rightmost). GT columns indicate ground truth images while in the OCC columns are presented the input occluded images. Columns 3 and 9 indicate the outputs of our baseline, where adversarial loss and MSE are used. 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.

two images. In fact, all the non-occluded parts that are visible in the input images must be transferred to the output with no alterations. The structure of the U-Net lends itself optimally to our task, and the skip connections are fundamental for the conservation of the non-occluded image content. In this way, useful low-level information is not lost during the encoding passage: by leveraging this kind of information, we are able to maintain the appearance of visible parts in the image.

Discriminator Network The Discriminator, instead, aims to determine if an image is true or if it has been generated. In particular, the structure is similar to the one proposed by [210], composed by 4 convolutional layers with kernel size 5×5. The resulting features are followed by one sigmoid activation function in order to obtain a probability for the classification problem. We use batch normalization before every Leaky ReLU activation, except for the first layer.

Table 3.16: Ablation study results on RAP dataset

Method	mAcc.	Acc.	Prec.	Rec.	F1	SSIM	PSNR
Occlusion	65.74	51.06	68.72	64.36	66.47	0.7153	14.57
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 + attr. loss	72.18	59.59	73.51	73.72	73.62	0.8239	20.65
VGG + attr. loss (+input)	**81.10**	**74.8**	**84.29**	**85.61**	**84.94**	0.8274	20.7
GT data	78,66	66,23	77.85	79.71	78.77	-	-

Table 3.17: Ablation study results on AiC dataset

Method	mAcc.	Acc.	Prec.	Rec.	F1	SSIM	PSNR
Occlusion	72.24	45.77	48.78	79.03	60.32	0.6148	18.38
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 + attr. loss	78.37	53.3	55.73	85.46	67.46	**0.7101**	**21.81**
VGG + attr. loss (+input)	**90.86**	**72.15**	**74.0**	**95.1**	**83.23**	0.6986	21.47
GT data	91.89	74.87	76.80	95.43	85.11	-	-

Training Details

We trained our GAN with 320×128 resized input images while simultaneously providing the target image in order to compute the supervised losses. We adopted the standard approach by [83] to optimize the networks alternating gradient descent updates between the Generator and the Discriminator with $K = 1$. Data augmentation is performed by randomly flipping the images horizontally. We used mini-batch SGD applying the Adam solver with momentum parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$, learning rate $2 \cdot 10^{-4}$ and a batch size of 20. Each training is performed using a Titan Xp GPU.

3.4.2 Datasets

We evaluated our method on the RAP dataset, proposed by [141], comparing state-of-the-art methods and performing the ablation study over each loss employed. In addition, we further propose a new large-scale computer-graphics dataset AiC

for pedestrian attribute recognition in crowded scenes. Differently, from existing publicly available datasets, AiC is mainly focused on occlusion events.

RAP Dataset

RAP by [141] is a very richly annotated dataset with 41,585 pedestrian samples, each of which is labeled with 72 attributes as well as viewpoints, occlusions, and body parts information. In order to evaluate our method, we corrupted the dataset with occlusions. Differently, from what did by [60], where obstructions were created by cutting parts of images according to regular geometric shapes, we have adopted a more sophisticated approach that has led us to more realistic results. By exploiting the state-of-the-art performances of Mask R-CNN proposed by [93], pre-trained on the COCO Dataset ([149]), we produced segmentation masks for each person in the RAP dataset. The computed silhouettes were then used to crop people's shapes from the dataset. Those figures are then used to reproduce the occlusions, simply by randomly overlapping the crops to each image sample of RAP dataset. In addition, to reduce the visual gap between the original image and the overlapped person, we performed a Gaussian blurring. However, this is not applied to the whole image but only to the area given by the difference between an expansion and an erosion of the segmentation mask of the overlapping image. The only constraint that we have introduced is that the occluding person must not occupy the portion of the image that has the y coordinate that exceeds the 6/7 of the image height. Each sample is computed as follows:

$$I_{OCC} = I_{GT^1} \odot \neg\alpha\left(\beta\left(I_{GT^2}\right)\right) + \alpha\left(\beta\left(I_{GT^2}\right) \odot I_{GT^2}\right) \tag{3.33}$$

where $\beta(I_{GT^2})$ is the binary mask generated using Mask R-CNN and morphology operations and α is a function used to translate the overlap section randomly over the destination image I_{GT^1}. The dataset is already organized in 5 random splits. Each of which contains 33,268 images for training and 8,317 for testing. As did by [141], due to the unbalanced distribution of attributes in RAP, we selected the 51 attributes that have the positive example ratio in the dataset higher than 0.01.

AiC Dataset

Most of the publicly available pedestrian attribute datasets, like RAP by [141], PETA by [50] and PA-100K by [153] does not contemplate occlusion events. They only provide samples of full visible people, completely ignoring crowded situations of pedestrians occluding each other (which is indeed common in urban

scenarios). To overcome this limitation, we propose the Attributes in Crowd dataset, a novel synthetic dataset for people attribute recognition in presence of strong occlusions. AiC features 125,000 samples, all being a unique person, each of which is automatically labeled with information concerning sex, age etc. The dataset is split into 100,000 samples for training and 25,000 for testing purposes. Each of the 24 attributes is present at least in a 10% of samples which highlight a good balance in terms of labels. The collected samples feature a vast number of different body poses, in several urban scenarios with varying illumination conditions and viewpoints. Skeleton joints are also available for each identity. Joints are additionally labeled with an occlusion flag which tells if the specific body part is directly visible from the camera point of view. Moreover, each image sample has his 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 obviously not achievable in real environments). Fig. 3.26 exhibits some examples of the dataset. AiC was created by exploiting the highly photo-realistic video game *Grand Theft Auto V* developed by *Rockstar North*. To foster future research on this topic, the dataset is publicly available here `https://github.com/fabbrimatteo/AiC-Dataset`.

3.4.3 Experiments

In this section, we provide details about the metrics adopted, followed by a detailed ablation study that presents qualitative and quantitative results for three different combinations of losses (that we added to the adversarial loss): MSE loss, VGG loss and a combination of VGG loss and attribute loss. We also investigate how the information about the attributes of a person can enhance the quality of the produced images. Additionally, we explain the choice of different hyperparameters, exploring their impacts. Finally, we compare our method with the most related works presented by [108] and [60].

Evaluation Metrics

Evaluating the quality of synthesized images is an open and challenging problem as stated by [233]. Traditional metrics such as per-pixel MSE do not estimate joint statistics of the result, and therefore do not extrapolate the full structure of the result. In order to more holistically evaluate the visual quality of our results, we employed two tactics. Firstly, we compared the performance of the proposed model through metrics directly calculated over the reconstructed images. Specifically, we

Figure 3.28: Qualitative results on both RAP and AiC datasets. (first line) an example using a configuration of $\lambda_1 = 0$ and $\lambda_2 = 15$ on RAP: the color of the jacket mutate from pink to gray to facilitate the classification, as the majority of jackets in the dataset are dark. (second line) an example using a configuration of $\lambda_1 = 10$ and $\lambda_2 = 0.1$ on AiC: blurring the occlusion and not hallucinating new body parts results in a better strategy to facilitate ResNet-101.

adopted the structural similarity SSIM and the peak signal-to-noise ratio PSNR. Secondly, we measured the capability of the proposed network of being able to preserve original attributes, like gender, hairstyle or wearing jacket, by exploiting the ResNet-101 network of [95] trained on the task of multi-attribute classification. Thus, following [141], [60] and [153], we provide five evaluation metrics for the attribute classification task, namely mean Accuracy, Accuracy, Precision, Recall and F1.

ResNet-101 Classification Network We trained the network with 320×128 resized images with Adam as optimizer and learning rate set to $2 \cdot 10^{-4}$. In Table 3.14 a comparison on the classification task with other state-of-the-art networks on RAP dataset is presented. The same network is trained independently for each dataset, in order to provide reliable metrics for both RAP and AiC.

Ablation Study

As previously stated, we investigated three loss combinations in order to clarify and highlight the solutions adopted in our work:

- *Baseline*: the Baseline architecture uses, in conjunction with the adversarial loss, the MSE loss as content loss;

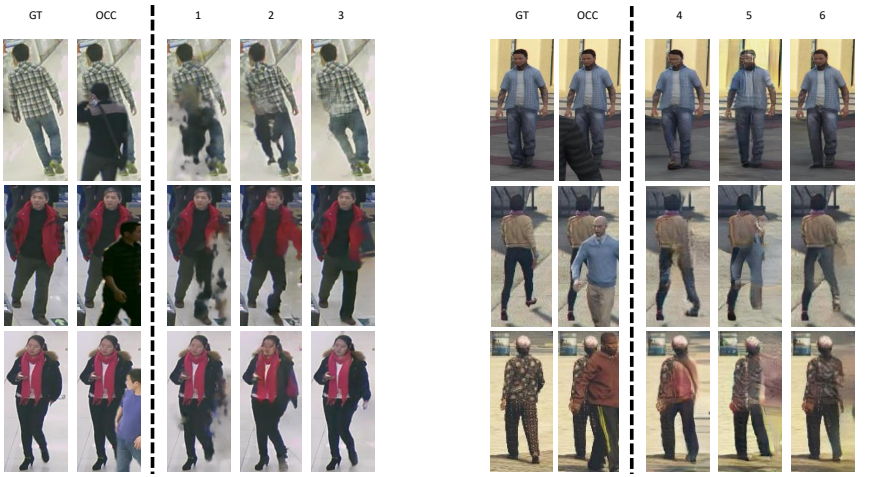

Figure 3.29: Qualitative comparison with state-of-the-art approaches: results are presented for both RAP (leftmost) and AiC (rightmost). GT columns indicate ground truth images while in the OCC columns are presented the input occluded images. Columns 1 and 4 are the images recovered by Pix2Pix. On 2 and 5 are presented results obtained from the method used by Fabbri et al. The last two columns, 3 and 5, show our best comparable approach output (Vgg loss + Attr. loss).

- *VGG loss*: differently from the Baseline, we replaced the MSE loss with the VGG loss. The layers (1,2), (2,2), (3,3) and (4,3) are chosen as the set L of activations on Eq. 3.31. In Eq. 3.28, we set λ_1 to 10 and λ_2 to 0 (further details about λ_1 and λ_2 are presented in the next subsection);

- *VGG loss + Attr. loss*: in this case, all the three losses are employed. The VGG loss always refers to the same four activation layers. The Attribute loss is computed between the output of the ResNet-101 classification network computed on the generated images and the ground truth labels provided by the datasets. In Eq. 3.28, we set λ_1 to 10 and λ_2 to 5 for RAP and to 0.01 for AiC. Note that we did not use all the available attributes of RAP dataset, but only the first 51 for the reason explained at the end of section 3.4.2. For AiC dataset, instead, we used all the available attributes.

In order to further investigate how some additional information about the attributes can improve the restoration process, we performed a further experiment where attributes are fed as input to the network, along with the occluded image:

- *Entire*: in this setup, we adopted both the VGG loss and the Attribute loss, along with the adversarial loss. Differently, from our main method, attributes are injected directly to the main flow of the Generator network. Specifically, the attribute vector of the occluded pedestrian is fed to a fully connected layer in order to produce a feature vector that is reshaped to match the bottleneck dimension of our Generator network.

Fig. 3.27 shows some qualitative results. The baseline performs considerably worse than the other setups, not being able to completely remove the occlusions on AiC (column 9 of Fig. 3.27). This is probably due to the fact that AiC is a more challenging dataset compared to our corrupted version of RAP. For the same reason, RAP results are overall more appealing than the ones of 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.

Table 3.16 and Table 3.17 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. By observing the tables, we can state that, 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

Table 3.18: Comparison with the state-of-the-art method on RAP dataset

Method	mAcc.	Acc.	Prec.	Rec.	F1	SSIM	PSNR
Occlusion	65.74	51.06	68.72	64.36	66.47	0.7153	14.57
Pix2Pix [108]	69.49	52.05	65.07	70.06	67.47	0.7348	17.91
[60]	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**

Table 3.19: Comparison with the state-of-the-art method on RAP dataset

Method	mAcc.	Acc.	Prec.	Rec.	F1	SSIM	PSNR
Occlusion	72.24	45.77	48.78	79.03	60.32	0.6148	18.38
Pix2Pix [108]	71.93	44.27	46.75	81.61	59.45	0.6351	21.22
[60]	67.14	38.21	40.61	79.9	53.85	0.573	20.11
Ours	**78.37**	**53.3**	**55.73**	**85.46**	**67.46**	**0.7101**	**21.81**

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, probably due to the more challenging nature of AiC. Moreover, Table 3.16 shows that the Entire setup reach higher scores compared to the upper bound of the ground truth images. Also Table 3.17 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). As can be shown in the next subsection, further forcing the generation output on classification metrics, we can reach results that exceed the ground truth upper bound even on AiC, at a price of a drop on reconstruction metrics.

Hyperparameter Optimization

Hyperparameter tuning is a crucial aspect in designing machine learning frameworks, as the performance of an algorithm can be highly dependent on the choice of hyperparameters. In fact, λ_1 and λ_2 were selected using a grid search technique. In particular, we searched for a trade off between classification metrics (accuracy,

precision, recall, f1) and pixel-level reconstruction metrics (PSNR, SSIM). We performed a different grid search for four different configurations combining each dataset with the two main setups: VGG loss + Attr. loss and the Entire pipeline.

For what concerns the VGG loss + Attr. loss setup, we observed that, in general, a configuration with $\lambda_1 \gg \lambda_2$ brings to better pixel-level reconstruction metrics but poor classification performances. On the other hand, solutions with $\lambda_1 \ll \lambda_2$ show good classification performances but low pixel-level reconstruction metrics. Also, increasing λ_1 over the value of 10 does not further improve PSNR and SSIM metrics (for both RAP and AiC). The same behavior happens for λ_2: the classification metrics do not improve for values greater than 5 (for RAP) and 0.01 (for AiC). This difference of λ_2 between the two datasets may be caused by the fact that AiC is more challenging than RAP. In fact, during training, the Attributes Loss on AiC is orders of magnitude greater than the same loss on RAP, thus, a smaller λ_2 is needed to maintain the balance between the losses.

For what concern the Entire pipeline, we observed a different behavior on λ_2: increasing λ_2 does steadily improve the classification metrics (reaching up to 98.89 mean Accuracy with $\lambda_2 = 5$) while drastically decreasing PSNR and SSIM. This behavior happens on both RAP and AiC. By giving more importance to the Attributes Loss, the Generator network is able to enhance attribute characteristics to the point that they are highly recognizable by the classification network, at the price of not maintaining low-level similarity. Fig. 3.28 shows a direct consequence at qualitative level on both RAP and AiC. The first line depicts an extreme configuration of $\lambda_1 = 0$ and $\lambda_2 = 15$ on RAP. With no low-level constraints, the Generator network is able to mutate the color of the jacket to facilitate the ResNet-101 "jacket attribute" recognition. The second line of Fig. 3.28, instead, shows an example obtained using $\lambda_1 = 10$ and $\lambda_2 = 0.1$ on AiC. In this case, the behavior is completely different: due to the high diversity of attributes in AiC, the Generator learns to simply remove the obstacle, not adding (hallucinating) many details to the removed portion of the image. Adding imprecise details would, in fact, mislead the attribute classification network.

Comparison Against State-Of-The-Art Techniques

Since our task of de-occlusion is novel, there are no direct works to compare with. So, to match the results of our network, in addition to our previous work, we also retrained the Pix2Pix framework on both RAP and AiC.

Our Previous work Like our current method, [60] exploits an adversarial based framework to achieve a translation from an occluded-pedestrian domain to a completely visible body domain. The main difference with our current method resides in the loss formulation: [60] minimizes a combination of adversarial loss and sum of squared error loss (SSE), completely ignoring high-level and low-level similarities. Another important difference lies in the Generator architecture: our previous work uses a simple hourglass architecture with no *skip connections*, while in our current method we adopted a U-net based solution. The U-net architecture shows better performances in tasks where some input information has to be shuttled directly to the output with no variation. In fact, as can be seen in Fig. 3.29, our previous work fails to preserve the portions of the image that should remain unchanged (especially the faces).

Pix2Pix [108] investigates conditional adversarial networks as a general-purpose solution to image-to-image translation problems. As in our Generator network, Pix2Pix exploits a U-net based architecture. The only substantial architectural difference is in the number of convolutional layers before each downsampling and after each upsampling operation. Also, the Discriminators differs: Pix2Pix uses a patch level discriminator that only penalizes structure at the scale of patches, while in our work we adopt an image level discriminator that takes the whole image as input. A patch level discriminator models the image as a Markov random field, assuming independence between pixels separated by more than a patch diameter. This is indeed not the case when dealing with images of people. In fact, for example, the skin color of the face should match the skin color of the hands. Also, the trousers are usually made of the same color. Another significant difference lies in the content loss: Pix2Pix, like our previous work, uses a pixel-level loss (L1 instead of SSE), assuming pixel independence, and forcing pixels of the output image to exactly match the pixels of the target image. In our work, instead, we exploit a combination of high-level and low-level consistency by encouraging the overall images the have similar feature representations as computed by the VGG16 network, and similar visual attributes as computed by the ResNet-101.

From Table 3.18 and Table 3.19 it can be shown that our network perform favourably for each metric, both for RAP and AiC datasets. From Fig. 3.29 it also emerges that our method, despite not using attention mechanisms, is able to detect and to remove the occlusion, with no external additional information. Furthermore, differently from the works by[60] and [108], our method learns to transfer with no alterations the portion of images that are not occluded.

Chapter 4

Domain Adaptation

Deep learning methods trained on synthetically generated data usually suffer from domain-shift related problems. For this reason, in this chapter, we investigate domain adaptation techniques in order to bridge the gap between "source domain" and "target domain" for head pose estimation, attributes recognition and facial landmark localization. In particular, in Section 4.1, we propose a complete framework for the estimation of the head and shoulder pose relying on depth images only where a Face-from-Depth component based on a Conditional GAN is able to hallucinate a face from the corresponding depth image. Additionally, in Section 4.2, we further explore the capabilities of the Face-from-Depth component. Although the network cannot reconstruct the exact somatic features for unknown individual faces, it is capable of reconstructing plausible faces as their appearance is accurate enough as it can be used as input for face attributes classification and land-mark localization.

4.1 Head Pose Estimation

Computer vision has been addressing the problem of head pose estimation for several years.
In 2009, Murphy-Chutorian and Trivedi [182] made a first assessment of the proposed techniques. More recently, different approaches have been proposed together with some annotated datasets useful for both training and testing those systems. The interest of the research community is mainly due to a large number

of applications that require or are improved by a reliable head pose estimation: face recognition with aliveness detection, human-computer interaction, people behavior understanding are some examples. Moreover, a large effort has been recently devoted to applications in the automotive field, such as monitoring drivers and passengers. Together with the estimation of the upper-body and shoulder pose, the head monitoring is one of the key technologies required to set up (semi)-autonomous driving cars, human-car interaction for entertainment, and driver's attention measurement.

In the automotive field, vision-based systems are required to cooperate or even replace other traditional sensors, due to the increasing presence of cameras inside new cars' cockpits and to the ease of capturing images and videos in a completely non-invasive manner.
In the past, encouraging results for driver head pose estimation have been achieved using RGB images [182, 258, 13, 54, 42] as well as different camera types, such as infrared [112], thermal [259], or depth [174, 161, 20]. Among them, the last ones are very promising, since they allow robustness when facing strong illumination variations. Moreover, standard techniques based on RGB images are not always feasible due to poor or absent illumination conditions during the night or to the continuous illumination changes during the day.

Nowadays, the acquisition of depth data is feasible thanks to commercial low-cost, high-quality and small-sized depth sensors, that can be easily placed inside the vehicle.

In this work, we propose a robust and fast solution for head and shoulder pose estimation, especially devoted to drivers in cars, but that can be easily generalized to any application where depth images are available. The presented framework provides impressive results, reaching an accuracy higher than 73% on the new *Pandora* dataset (see Fig. 4.3) and a low average error on the *Biwi* dataset, thus overcoming all state-of-art related works.

The core of the framework is a Convolutional Neural Network (CNN), called *POSEidon*$^+$, that combines depth, appearance and Motion Images as input to estimate the 3D pose angles in regression. An overview of the model is depicted in Figure 4.4. The model is enhanced with a *Face-from-Depth (FfD)* component. This is motivated by recent literature results[1, 55] that testifies the importance of intensity images for the task. The *FfD* component is able to reconstruct the gray-level appearance of a face directly from the corresponding depth image. Thanks to the insensitivity of the depth image to the external illumination conditions, the provided reconstruction is more stable and reliable than gray or color images captured from the same RGB-D sensor. Moreover, the reconstruction can be

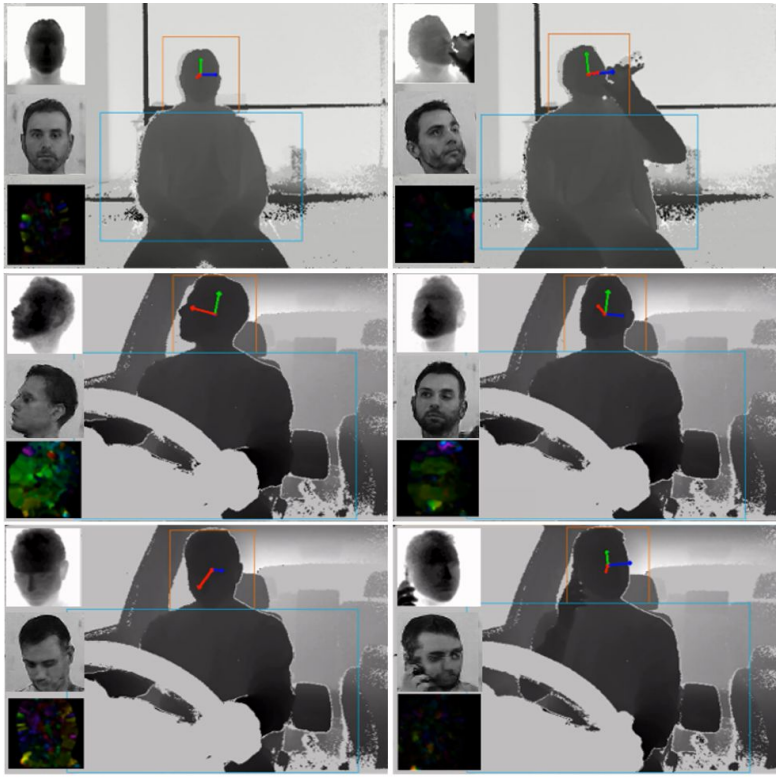

Figure 4.1: Visual examples of the proposed framework output in indoor (first row) and automotive (second and third row) settings. Head pose angles are reported as colored arrows. Depth maps, *Face-from-Depth* and Motion Image inputs are depicted on the left of each frame.

applied in situations where the depth sensor is exploited alone without the color stream for computational or implementation constraints.
As an example, in Figure 4.2, we have reported two frames captured from an RGB-D sensor in correspondence of an abrupt illumination change (from light to dark). The depth images are not affected by the illumination change and thus the corresponding *FfD* reconstructions are identical. The provided output highlights the reliability of the developed network as well as the quality of the results.

Figure 4.2: Example of reliability of the *FfD* network on depth images. Two consecutive frames have been selected from a sequence with an abrupt illumination change (from light to dark). In the first column the auto equalized RGB, then the corresponding depth maps and finally the *FfD* reconstruction output.

The overall system is split into two components: the *Face-from-Depth* architecture followed by the pose estimation module, that takes as input the reconstructed gray level images. From a first glance, this approach could be improper since we are somehow forcing the *FfD* model to output a human understandable intermediate representation, *i.e.*, the gray level image. Training an end-to-end system enables the network to find the best internal/intermediate representation. However, in addition to a performance improvement as reported in Section 4.1.4, the introduction of the *Face-from-Depth* component allows the second part of the system to be trained on wider datasets since more annotated datasets on gray-level images are usually available rather than on depth ones. More generally, *FfD* moves input depth images on a domain where more experience is available in order to understand and process them.

This work is an improved and extended version of our preliminary work, that has been described in [20], where the body pose estimation task was carried on through a baseline version of the *POSEidon*$^+$ framework, here referred as *POSEidon*. In this thesis we present the overall framework, introducing a new *Face-from-Depth* architecture, which exploits the recent *Deterministic Conditional GAN* models [107] to reconstruct gray-level face images. To the best of our knowledge, this is one of the first proposal to generate intensity images from depth data for the head pose estimation task with an *adversarial* approach. Moreover,

Figure 4.3: Sample frames from the *Pandora* dataset. As depicted, extreme poses and challenging camouflage can be present.

we evaluate and check the overall quality of the computed face images and results confirm their high quality and accuracy.
Extensive experiments have been carried out and results show that the *POSEidon*$^+$, equipped with the improved version of the *Face-from-Depth* architecture, achieves significant improvements in the head pose estimation task. Besides, we show that is possible to obtain competitive results exploiting a CNN trained on gray-level faces and tested on generated ones.

4.1.1 Method

An overview of the *POSEidon*$^+$ framework is depicted in Figure 4.4. The final goal is the estimation of the pose of the driver's head and shoulders, defined as the mass center position and the corresponding orientation relative to the reference frame of the acquisition device [182]. The orientation is provided using three rotation angles *pitch*, *roll* and *yaw*. *POSEidon*$^+$ directly processes the stream of depth frames captured in real-time by a commercial sensor. Position and size of the foremost head are estimated by a head localization module based on a regressive CNN (Sect. 4.1.1). The output provided is used to crop the input frames around the head and the shoulder bounding boxes, depending on the further pipeline type. Frames cropped around the head are fed to the head pose estimation block, while the others are exploited to estimate the shoulders pose.
The core components of the system are the *Face-from-Depth* network (Sect. 4.1.2), and *POSEidon*$^+$ (Sect. 4.1.3), the network which gives the name to the whole framework. Its trident shape is due to the three included CNNs, each working on a specific source: depth, gray level (the output of *FfD*) and *Motion Images* data. The first one plays the main role on the pose estimation, while the others cooperate to reduce the estimation error.

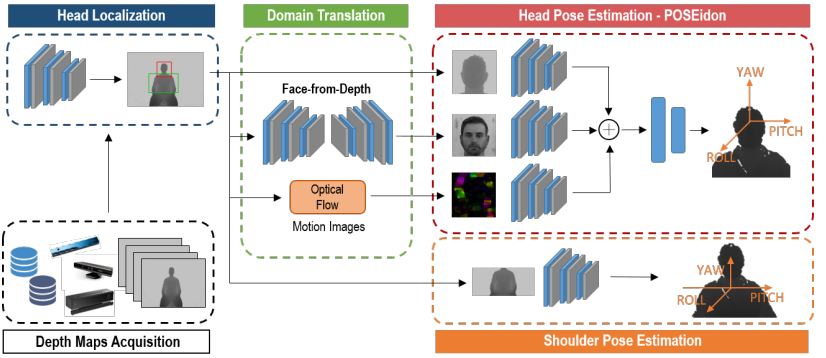

Figure 4.4: Overview of the whole *POSEidon*$^+$ framework. Depth input images are acquired by depth sensors (black) and provided to a head localization CNN (blue) to suitably crop the images around the upper-body or head regions. The head crop is used to produce the three inputs for the following networks (green), that are then merged to output the head pose (red). In particular, the *Face-from-Depth* architecture reconstructs gray-level face images from the corresponding depth maps, while the Motion Images are obtained by applying the *Farneback* algorithm. Finally, the upper-body crop is used for the shoulder pose estimation (orange). [best in color]

Head Localization

In this step, we defined and trained a network for head localization, relying on the main assumption that a single person is in the foreground. The image coordinates (x_H, y_H) of the head center are the network outputs, or rather, the center mass position of all head points in the frame [243].
Details on the deep architecture adopted are reported in Figure 4.5. A limited depth and small-sized filters have been chosen to meet real-time constraints while keeping satisfactory performance. For this reason, input images are firstly resized to 160×132 pixels. A max-pooling layer (2×2) is run after each of the first four convolutional layers, while a dropout regularization ($\sigma = 0.5$) is exploited in fully connected layers. The hyperbolic tangent activation (*tanh*) function is adopted, in order to map continuous output values to a predefined range $[-\infty, +\infty] \rightarrow [-1, +1]$. The network has been trained by *Stochastic Gradient Descent* (SGD) [128] and the L_2 loss function.

Given the head position (x_H, y_H) in the frame, a dynamic size algorithm provides the head bounding box with center (x_H, y_H), width w_H and height h_H, around which the frames are cropped:

$$w_H = \frac{f_x \cdot R_x}{D}, \quad h_H = \frac{f_y \cdot R_y}{D} \tag{4.1}$$

where f_x, f_y are the horizontal and the vertical focal lengths in pixels of the acquisition device, respectively. R_x, R_y are the average width and height of a face (for head pose task $R_x = R_y = 320$) and D is the distance between the head center and the acquisition device, computed averaging the depth values around the head center.
Some examples of bounding boxes estimated by the network are superimposed in the frames of Figure 4.1.

4.1.2 Face-from-Depth

Due to illumination issues, the appearance of the face is not always available if acquired with a RGB camera, *e.g.* inside a vehicle at night. On the contrary, depth maps are generally invariant to illumination conditions but lack of texture details. We aim to investigate if it is possible to imagine the appearance of a face given

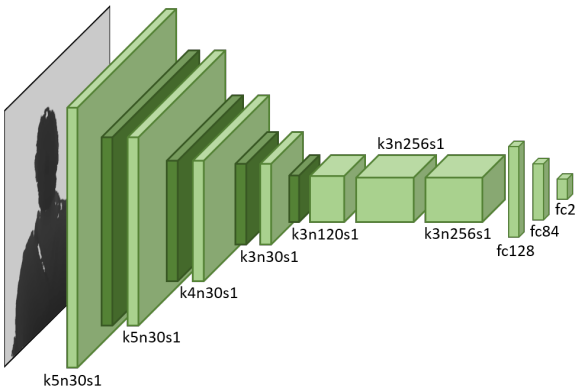

Figure 4.5: Architecture of the Head Localization network with corresponding kernel size (k), number of feature maps (n) and stride (s) indicated for each convolutional layer.

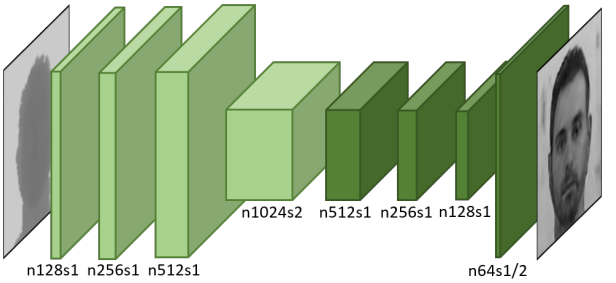

Figure 4.6: Architecture of the *Face-from-Depth* network.

the corresponding depth data. Metaphorically, we ask the model to mimic the behavior of a blind person when he tries to figure out the appearance of a friend through the touch.

Deterministic Conditional GAN

The *Face-From-Depth* network exploits the *Deterministic Conditional GAN* (detcGAN) paradigm [107] and it is obtained as a generative network G capable of estimating a gray-level image I^E of a face from the corresponding depth represent-

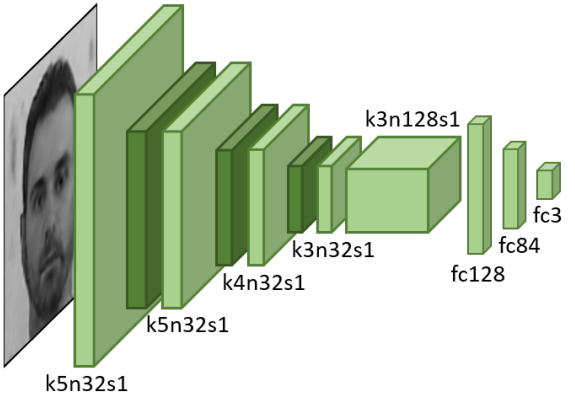

Figure 4.7: Architecture of the Head and Shoulder Pose Estimation networks.

Table 4.1: Head Pose Estimation Results on *Biwi*. To allow fair comparisons with state of the art methods, **POSEidon**$^+$ has been evaluated using different evaluation protocols.

Validation Procedure	Year	Data		Head			Avg
		Depth	RGB	Pitch	Roll	Yaw	
ALL SEQUENCES USED AS TEST SET							
Padeleris [195]	2012	√		6.6	6.7	11.1	8.1
Rekik [216]	2013	√	√	4.3	5.2	5.1	4.9
Martin [164]	2014	√		2.5	2.6	3.6	2.9
Papazov [199]	2015	√		2.5 ± 7.4	3.8 ± 16.0	3.0 ± 9.6	4.0 ± 11.0
Meyer [174]	2015	√		2.4	2.1	2.1	2.2
Li [142]	2016	√	√	1.7	3.2	2.2	2.4
Sheng [241]	2017	√		2.0	1.9	2.3	2.1
LEAVE ONE OUT (LOO)							
Drouard [55]	2015		√	5.9 ± 4.8	4.7 ± 4.6	4.9 ± 4.1	5.2 ± 4.5
Drouard [56]	2017		√	10.0 ± 8.7	8.4 ± 8.0	8.6 ± 7.2	9.0 ± 7.9
POSEidon$^+$	2017	√		$\mathbf{2.4 \pm 1.3}$	$\mathbf{2.6 \pm 1.5}$	$\mathbf{2.9 \pm 1.5}$	$\mathbf{2.6 \pm 1.4}$
K4-FOLD SUBJECT CROSS VALIDATION							
Fanelli [64]	2011	√		3.5 ± 5.8	5.4 ± 6.0	3.8 ± 6.5	- ± -
POSEidon$^+$	2017	√		$\mathbf{2.8 \pm 1.7}$	$\mathbf{2.9 \pm 2.1}$	$\mathbf{3.6 \pm 2.5}$	$\mathbf{3.1 \pm 2.1}$
K5-FOLD SUBJECT CROSS VALIDATION							
Fanelli [66]	2011	√		8.5 ± 9.9	7.9 ± 8.3	8.9 ± 13.0	8.43 ± 10.4
POSEidon$^+$	2017	√		$\mathbf{2.8 \pm 1.8}$	$\mathbf{2.8 \pm 2.2}$	$\mathbf{3.6 \pm 2.2}$	$\mathbf{3.0 \pm 2.1}$
K8-FOLD SUBJECT CROSS VALIDATION							
Lathuiliere [133]	2017		√	4.7	3.1	**3.1**	3.6
POSEidon$^+$	2017	√		$\mathbf{2.8 \pm 1.9}$	$\mathbf{2.8 \pm 1.8}$	3.3 ± 2.0	$\mathbf{3.0 \pm 1.9}$
FIXED TRAIN AND TEST SPLITS							
Yang [279]	2012	√	√	9.1 ± 7.4	7.4 ± 4.9	8.9 ± 8.3	8.5 ± 6.9
Baltrusaitis [11]	2012	√	√	5.1	11.3	6.3	7.6
Kaymak [119]	2013	√	√	7.4	6.6	5.0	6.3
Wang [263]	2013	√	√	8.5 ± 14.3	7.4 ± 10.8	8.8 ± 14.3	8.2 ± 12.0
Ahn [1]	2014		√	3.4 ± 2.9	2.6 ± 2.5	2.8 ± 2.4	2.9 ± 2.6
Saeed [232]	2015	√	√	5.0 ± 5.8	4.3 ± 4.6	3.9 ± 4.2	4.4 ± 4.9
Liu [152]	2016		√	6.0 ± 5.8	5.7 ± 7.3	6.1 ± 5.2	5.9 ± 6.1
POSEidon [20]	2017	√		1.6 ± 1.7	1.8 ± 1.8	1.7 ± 1.5	1.7 ± 1.7
POSEidon$^+$	2017	√		$\mathbf{1.6 \pm 1.3}$	$\mathbf{1.7 \pm 1.7}$	$\mathbf{1.7 \pm 1.3}$	$\mathbf{1.6 \pm 1.4}$

ation I^D. The generator G is trained to produce outputs as much indistinguishable as possible from *real* images I by an adversarially trained discriminator D, which is expressly trained to distinguish the *real* images from the *fake* ones produced by the generator. Differently from a traditional GAN [84, 210], the Generator network of a det-cGAN takes an image as input (to be *Conditional*) and not a random noise vector (to be *Deterministic*). As a result, a det-cGAN learns a mapping from observed images x to output images y: $G : x \rightarrow y$.

The objective of a det-cGAN can be expressed as follows:

$$L_{det-cGAN}(G, D) = \mathbb{E}_{I \sim p_{data}(I)}[\log D(I)] + \mathbb{E}_{I^D \sim p_{data}(I^D)}[\log(1 - D(G(I^D)))] \quad (4.2)$$

where $\log D(I)$ represents the log probability that I is *real* rather than *fake* while $\log(1 - D(G(I^D)))$ is the log probability that $G(I^D)$ is *fake* rather than *real*. G tries to minimize the term $L_{det-cGAN}(G, D)$ of Equation 4.2, against D that tries to maximize it. The optimal solution is:

$$G^* = \arg \min_G \max_D L_{det-cGAN}(G, D) \quad (4.3)$$

As a possible drawback, the images generated by G are forced to be realistic thanks to D, but they can be unrelated with the original input. For instance, the output could be a nice image of a head with a very different pose with respect to the input depth. Thus, is fundamental mixing the GAN objective with a more traditional loss, such as SSE distance [202]. While discriminator's job remains unchanged, the generator, in addition to fooling the discriminator, tries to emulate the ground truth output in an SSE sense. The pixel-wise SSE is calculated between downsized versions of the generated and target images, first applying an averaged pooling layer. We formulate the final objective as the weighted sum of a content loss and an adversarial loss as:

$$G^* = \arg \min_G \max_D L_{det-cGAN}(G, D) + \lambda L_{SSE}(G) \quad (4.4)$$

where λ is the weight controlling the content loss impact.

Network Architecture

We propose to modify the classic hourglass generator architecture, performing a limited number of upsampling and downsampling operations. As shown in the

Table 4.2: Evaluation metrics computed on the reconstructed gray-level face images with *Biwi* and *Pandora* datasets. Starting from the left, L_1 and L_2 distances are reported, then the absolute and the squared differences, the root-mean-square error and, finally, the percentage of pixels under a certain threshold.

Dataset	Method	Norm ↓		Difference ↓		RMSE ↓			Threshold ↑		
		L_1	L_2	Abs	Squared	linear	log	scale-inv	1.25	2.5	3.75
Biwi	FfD [20]	33.35	2586	0.454	24.07	40.55	**0.489**	**0.445**	0.507	**0.806**	**0.878**
	FfD	**24.44**	**2230**	**0.388**	**19.81**	**35.50**	0.653	0.610	**0.615**	0.764	0.840
Pandora	FfD [20]	41.36	3226	0.705	46.00	50.77	0.603	**0.485**	0.263	0.725	0.819
	pix2pix [107]	19.37	1909	**0.468**	24.07	30.80	0.568	0.539	0.583	0.722	0.813
	AVSS [61]	23.93	2226	0.629	34.49	35.46	0.658	0.579	0.541	0.675	0.764
	FfD + U-Net	23.75	2123	0.653	34.96	33.89	0.639	0.553	0.555	0.689	0.775
	FfD	**18.21**	**1808**	0.469	**22.90**	**28.90**	**0.556**	0.501	**0.605**	**0.743**	**0.828**

following experimental section, the *U-Net* architecture [228] can be adopted in order to shuttle low-level information between input and output directly across the network [107], but it is less convenient in our case.
Following the main architecture guidelines for stable Deep Convolutional GANs by Radford *et al.* [210], we instead adopt the architecture illustrated in Figure 4.6 for the Generator. Specifically, in the encoder part, we use three convolutional layers followed by a strided convolutional layer (with stride 2) to halve the image resolution.
The decoding stack uses three convolutional layers followed by a transposed convolutional layer (also referred as fractionally strided convolutional layers) with stride $1/2$ to double the resolution, and a final convolution. The number of filters follows a power of 2 pattern, from 128 to 1024 in the encoder and from 512 to 64 in the decoder. *Leaky ReLU* is used as activation function in the encoding phase while *ReLU* is used in the decoding phase.
We adopt *batch normalization* before each activation (except for the last layer) and a kernel size 5×5 for each convolution.
The discriminator architecture complies with the generator's encoder in terms of activation and number of filters, but contains only strided convolutional layers (with stride 2) to halve the image resolution each time the number of filters is doubled. The network then outputs one *sigmoid* activation. In the discriminator, we use batch normalization before every Leaky ReLU activation, except for the first layer.

Training details

We trained the det-cGAN with depth images and simultaneously providing the network with the original gray-level images associated with the depth data in order to compute the L_{SSE}. To optimize the network we adopted the standard approach from Goodfellow *et al.* [84] and alternate the gradient descent updates between the generator and the discriminator with $K = 1$. We used mini-batch SGD applying the *Adam* solver [124] with $\beta_1 = 0.5$ and batch size of 64. We set $\lambda = 10^{-1}$ in Equation 4.4 for the experiments. Moreover, to encourage the discriminator to estimate soft probabilities rather than to extrapolate extremely confident classifications, we used a technique called *one-sided label smoothing* [233] where the target for the real examples are replaced with a value slightly less than 1, such as 0.9. This solution prevents the discriminator to produce extremely confident predictions that could unbalance the adversarial learning.

4.1.3 Pose Estimation from depth

POSEidon$^+$ network

The *POSEidon*$^+$ network is a fusion of three CNNs and has been developed to perform a regression on the 3D pose angles. As a result, continuous Euler values – corresponding to the *yaw*, *pitch* and *roll* angles – are estimated (right part of Fig. 4.4). The three *POSEidon*$^+$ components have the same shallow architecture based on 5 convolutional layers with kernel size of 5×5, 4×4 and 3×3 and a 2×2 max-pooling is conducted only on the first three layers due to the limited size of the input (64×64). The first four convolutional layers have 32 filters each, the last one has 128 filters. *tanh* is exploited as activation function; we are aware that *ReLU* [183] converges faster, but better performance in term of accuracy prediction are achieved.
The three networks are fed with different input data types: the first one, directly takes as input the head-cropped depth images; the second one is connected to the *Face-from-Depth* output and the last one operates on Motion Images, obtained applying the standard *Farneback* algorithm [67] on pairs of consecutive depth frames. The presence of depth discontinuities around the nose and the eyes generates specific motion patterns which are related to the head pose. Motion Images, thus, provide useful information for the estimation of the pose of a moving head. Frames with motionless heads are very rare in real videos. However, in those cases the common image compression creates artifacts around the face landmarks which allow the estimation of the head pose.

A fusion step combines the contributions of the three above described networks. The last fully connected layer of each component is removed in order to provide the following layers with more data and not only the estimated angles. As a results, the output of the whole *POSEidon*$^+$ network is not only a weighted mean of the three component outputs, but a more complex combination. Different fusion approaches that have been proposed by Park *et al.* [200] are investigated. Given two feature maps x^a, x^b with a certain width w and height h, for every feature channel d_a^x, d_b^x and $y \in R^{w \times h \times d}$:

- **Multiplication**: computes the element-wise product of two feature maps, as $y^{mul} = x^a \circ x^b, d^y = d_a^x = d_b^x$
- **Concatenation**: stacks two features maps, without any blend $y^{cat} = [x^a | x^b], d^y = d_a^x + d_b^x$
- **Convolution**: stacks and convolves feature maps with a filter k of size $1 \times 1 \times (d_a^x + d_b^x)/2$ and β as bias term, $y^{conv} = y^{cat} * k + \beta, \quad d^y = (d_a^x + d_b^x)/2$

The final *POSEidon*$^+$ framework exploits a combination of two fusing methods, in particular, a convolution followed by a concatenation. After the fusion step, three fully connected layers composed of 128, 84 and 3 activations respectively and two dropout regularization ($\sigma = 0.5$) complete the architecture. *POSEidon*$^+$ is trained with a double-step procedure. First, each individual network described above is trained with the following L_2^w weighted loss:

$$L_2^w = \sum_{i=1}^{3} \left\| w_i \cdot (y_i - f(x_i)) \right\|_2 \tag{4.5}$$

where $w_i \in [0.2, 0.35, 0.45]$. This weight distribution gives more importance to the yaw angle, which is preponderant in the selected automotive context. During the individual training step, the last fully connected layer of each network is preserved, then is removed to perform the second training phase. Holding the weights learned for the trident components, the new training phase is carried out on the last three fully connected layers of *POSEidon*$^+$ only, with the loss function L_2^w reported in Equation 4.5. In all training steps, the SGD optimizer [128] is exploited, the learning rate is set initially to 10^{-1} and then is reduced by a factor 2 every 15 epochs.

Shoulder Pose Estimation

The framework is completed with an additional network for the estimation of the shoulder pose. We employ the same architecture adopted for the head (Sect. 4.1.3), performing regression on the three pose angles.
Starting from the head center position (Sect. 4.1.1), the depth input images are cropped around the driver neck, using a bounding box $\{x_S, y_S, w_S, h_S\}$ with center $(x_S = x_H, y_S = y_H - (h_H/4))$, and width and height obtained as described in Equation 4.1, but with different values of R_x, R_y to produce a rectangular crop: these values are tested and discussed in Section 4.1.4. The network is trained with SGD optimizer [128], using the weighted L_2^w loss function described above (see Eq. 4.5). As usual, hyperbolic tangent is exploited as activation function.

4.1.4 Experiments

Datasets

Network training and testing phases have been done exploiting two publicly available datasets, namely *Biwi Kinect Head Pose* and *ICT-3DHP*. In addition, we collected a new dataset, called *Pandora*, which also contains shoulder pose annotations. Data augmentation techniques are employed to enlarge the training set, in order to achieve space invariance and avoid overfitting [128].
Random translations on vertical, horizontal and diagonal directions, jittering, zoom-in and zoom-out transformation of the original images have been exploited. Percentile-based contrast stretching, normalization and scaling of the input images are also applied to produce zero mean and unit variance data.
Other datasets for head pose estimation and related tasks have been collected in last decades [9, 85, 190, 289, 165], but in most cases there are some not desirable features, for instance, no depth or 3D data, no continuous ground truth annotations and not enough data for deep learning techniques.
Follows a detailed description of the three adopted datasets.

Biwi Kinect Head Pose dataset Fanelli *et al.* [65] introduced this dataset in 2013. It is acquired with the *Microsoft Kinect* sensor, i.e., a structured IR light device. It contains about 15k frames, with RGB (640×480) and depth maps (640×480). Twenty subjects have been involved in the recordings: four of them were recorded twice, for a total of 24 sequences. The ground truth of yaw, pitch and roll angles is reported together with the head center and the calibration matrix.

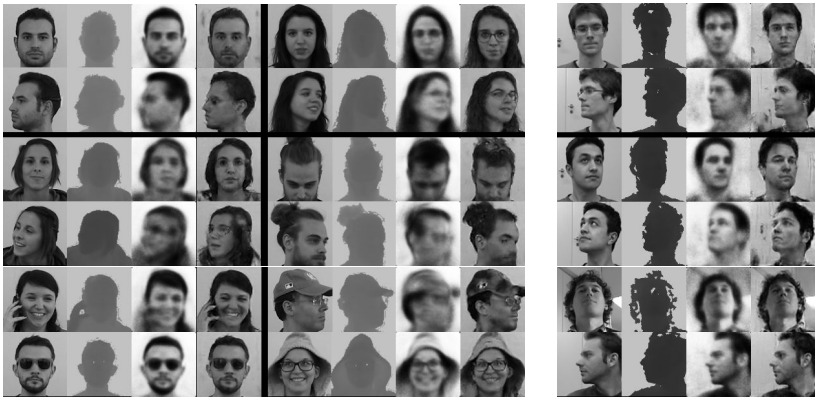

Figure 4.8: Test (a) and train (c) images on *Pandora* dataset, test (b) and train (d) images on *Biwi* dataset. For each block, gray-level images and then the corresponding depth faces are depicted in the first columns; face images taken from the method described by Borghi et al. are reported in the third column; finally, the output of the *Face-from-Depth* network proposed in this work is depicted in the last column.

The original paper does not report the adopted split between training and testing sets; fair comparisons are thus not guaranteed. To avoid this, we clearly report the adopted split in the following.

ICT-3DHP dataset *ICT-3DHP* dataset has been introduced by Baltrusaitis *et al.* in 2012 [11]. It has been collected using a *Microsoft Kinect* sensor and contains RGB images and depth maps of about 14k frames, divided into 10 sequences. The image resolution is 640×480 pixels. An additional hardware sensor (*Polhemus Fastrack*) is exploited to generate the ground truth annotation. The device is placed on a white cap worn by each subject, visible in both RGB and depth frames. The presence of a few subjects and the limited number of frames make this dataset unsuitable for training deep learning approaches.

Pandora dataset In addition to publicly available datasets, we have also collected and used a new challenging dataset, called *Pandora*. It has been specifically created for head center localization, head pose and shoulder pose estimation in automotive contexts (See Fig.4.3). A frontal and fixed device acquires the upper

body part of the subjects, simulating the point of view of a camera placed inside the dashboard. The subjects mainly perform driving-like actions, such as holding the steering wheel, looking to the rear-view or lateral mirrors, shifting gears and so on. *Pandora* contains 110 annotated sequences of 10 male and 12 female actors. Each subject has been recorded five times. *Pandora* is the first publicly available dataset which combines the following features:

- **Shoulder angles**: in addition to the head pose annotation, *Pandora* contains the ground truth data of the shoulder pose expressed as yaw, pitch, and roll.
- **Wide angle ranges**: subjects perform wide head ($\pm 70°$ roll, $\pm 100°$ pitch and $\pm 125°$ yaw) and shoulder ($\pm 70°$ roll, $\pm 60°$ pitch and $\pm 60°$ yaw) movements. For each subject, two sequences are performed with constrained movements, changing the yaw, pitch and roll angles separately, while three additional sequences are completely unconstrained.
- **Challenging camouflage**: garments, as well as various objects are worn or used by the subjects to create head and/or shoulder occlusions. For example, people wear prescription glasses, sunglasses, scarves, caps, and manipulate smart-phones, tablets or plastic bottles.
- **Deep-learning oriented**: the dataset contains more than 250k full resolution RGB (1920×1080) and depth images (512×424) with the corresponding annotation.
- **Time-of-Flight (ToF) data**: a *Microsoft Kinect One* device is used to acquire depth data, with a better quality than other datasets created with the first *Kinect* version [237].

Each frame of the dataset is composed of an RGB appearance image, the corresponding depth map, and the 3D coordinates of the skeleton joints corresponding to the upper body part, including the head center and the shoulder positions. For convenience's sake, the 2D coordinates of the joints on both color and depth frames are provided as well as the head and shoulder pose angles with respect to the camera reference frame. Shoulder angles are obtained through the conversion to Euler angles of a corresponding rotation matrix, obtained from a user-centered system [197] and defined by the following unit vectors (N_1, N_2, N_3):

$$\begin{aligned} N_1 &= \frac{p_{RS} - p_{LS}}{\|p_{RS} - p_{LS}\|} \quad U = \frac{p_{RS} - p_{SB}}{\|p_{RS} - p_{SB}\|} \\ N_3 &= \frac{N_1 \times U}{\|N_1 \times U\|} \qquad N_2 = N_1 \times N_3 \end{aligned} \tag{4.6}$$

Table 4.3: Results obtained on *Pandora* dataset with head pose network trained on gray level images and tested with the original gray-level and reconstructed ones.

Testing input	**Head**			**Acc.**
	Pitch	Roll	Yaw	
gray-level	**7.1 ± 5.6**	**5.6 ± 5.8**	**9.0 ± 10.9**	**0.613**
pix2pix [107]	7.9 ± 8.0	5.9 ± 6.3	12.8 ± 21.4	0.581
AVSS [61]	8.9 ± 8.5	6.2 ± 6.4	13.4 ± 20.4	0.543
FfD [20]	8.5 ± 8.9	6.1 ± 6.2	12.4 ± 17.3	0.559
FfD + U-Net	8.7 ± 8.4	6.4 ± 6.6	13.5 ± 19.9	0.552
FfD	**7.6 ± 6.9**	**5.8 ± 6.0**	**10.1 ± 12.6**	**0.613**

where p_{LS}, p_{RS} and p_{SB} are the 3D coordinates of the left shoulder, right shoulder and spine base joints. The annotation of the head pose angles has been collected using a wearable *Inertial Measurement Unit* (IMU) sensor. To avoid distracting artifacts on both color and depth images, the sensor has been placed in a non-visible position, *i.e.*, on the rear of the subject's head. The IMU sensor has been calibrated and aligned at the beginning of each sequence, assuring the reliability of the provided angles. The dataset is publicly available (`http://aimagelab.ing.unimore.it/pandora/`).

Table 4.4: Results of the head pose estimation on *Pandora* comparing different system architectures. The baseline is a single CNN working on the source depth map (Row 1). The accuracy is the percentage of correct estimations (err <15°). FfD: *Face-from-Depth*, MI: *Motion Images*.

HEAD POSE ESTIMATION ERROR [EULER ANGLES]

#	**Input**				**Crop**	**Fusion**	**Head**			**Accuracy**
	Depth	FfD	MI	Gray			Pitch	Roll	Yaw	
1	√					-	8.1 ± 7.1	6.2 ± 6.3	11.7 ± 12.2	0.553
2	√				√	-	6.5 ± 6.6	5.4 ± 5.1	10.4 ± 11.8	0.646
3		√			√	-	6.8 ± 6.1	5.8 ± 5.0	10.1 ± 12.6	0.658
4			√		√	-	7.7 ± 7.5	5.3 ± 5.7	10.0 ± 12.5	0.609
5				√	√	-	7.1 ± 6.6	5.6 ± 5.8	9.0 ± 10.9	0.639
6	√	√			√	concat	5.6 ± 5.0	4.9 ± 5.0	9.7 ± 12.1	0.698
7	√		√		√	concat	6.0 ± 6.1	4.5 ± 4.8	9.2 ±11.5	0.690
8	√	√	√		√	conv+concat	**5.6 ± 5.2**	**4.8 ± 5.0**	**8.2 ± 9.8**	**0.736**

Quantitative Results

The proposed framework has been deeply tested using the datasets described in Section 4.1.4. For evaluation with the *Pandora* dataset, sequences of subjects 10, 14, 16 and 20 have been used for testing, the remaining for training and validation. With *Biwi* dataset, test subjects are determined by the validation procedure adopted. Finally, we tested the system on all the sequences contained in *ICT-3DHP* dataset.

Domain Translation. First, we check the capabilities of the *Face-from-Depth* network alone. Some visual examples of input, output, and ground-truth images are reported in Figure 4.8.
With this aim, we propose two types of evaluation. The first is based on metrics related to the reconstruction accuracy. Following the work of Eigen etal [58], Table 4.2 reports some results. The system is evaluated both on *Biwi* and on *Pandora* datasets. *FfD* network is compared with other Image-to-Image methods taken from the recent literature. In particular, we trained from scratch the deep models proposed in [107, 61] (referred here as *pix2pix* and *AVSS*, respectively), following procedures reported in the corresponding papers.
Moreover, in order to investigate how architectural choices impact the reconstruction quality of *FfD*, we tested a different design. We modified the network adding the *U-Net* [228] skip connections between mirrored layers (cf. Sect. 4.1.2).
We also compared the presented approach with our preliminary version of *Face-from-Depth* network [20], that fuses the key aspects of *encoder-decoder* [167] and *fully convolutional* [155] neural networks.
For the sake of comparison, we report here key details about the preliminary *FfD* version [20]. It has been trained in a single step, with input head images resized to 64×64 pixels. The activation function is the hyperbolic tangent and best training performance are reached through the self adaptive *Adadelta* optimizer [292]. A particular loss function is exploited in order to highlight the central area of the image, where the face is supposed to be after the cropping step, and takes in account the distance between the reconstructed image and the corresponding gray-level ground truth:

$$L = \frac{1}{R \cdot C} \sum_{i}^{R} \sum_{j}^{C} \left(||y_{ij} - \bar{y}_{ij}||_2^2 \cdot w_{ij}^{\mathcal{N}} \right) \tag{4.7}$$

where R, C are the number of rows and columns of the input images, respectively. $y_{ij}, \bar{y}_{ij} \in \mathcal{R}^{ch}$ are the intensity values from ground truth ($ch = 1$)

Table 4.5: Results for head pose estimation task on *Pandora* dataset. In particular, here we compare our preliminary work with the proposed one. In addition, we include a comparison with *POSEidon*$^+$ framework, in which we replace the head pose estimation network trained on reconstructed face images with the same network trained on gray-level images, here referred as *POSEidon**.

Method	**Head**			**Acc.**
	Pitch	Roll	Yaw	
POSEidon [20]	5.7 ± 5.6	4.9 ± 5.1	9.0 ± 11.9	0.715
POSEidon*	5.6 ± 5.8	4.8 ± 5.0	8.8 ± 10.9	0.720
POSEidon$^+$	$\mathbf{5.6 \pm 5.2}$	$\mathbf{4.8 \pm 5.0}$	$\mathbf{8.2 \pm 9.8}$	**0.736**

and predicted appearance images. Finally, the term $w_{ij}^{\mathcal{N}}$ introduces a bivariate Gaussian prior mask. Best results have been obtained using $\mu = [\frac{R}{2}, \frac{C}{2}]^T$ and $\Sigma = \mathbb{I} \cdot [(R/\alpha)^2, (C/\beta)^2]^T$ with α and β empirically set to $3.5, 2.5$ for squared images of $R = C = 64$. Other details about network architecture and training are reported in [20].

The second set of tests is specific to the head pose estimation task. The head pose network described in Section 4.1.3, trained with gray-level images taken from the *Pandora* dataset, is tested on the reconstructed face images. Since the network has been trained on real gray-level images to output the angles of the head pose, we can suppose that the more generated images are similar to the corresponding gray-level ones, the better the results are. The comparison is presented in Table 4.3. In the first row, results obtained using gray-level images as testing input are reported, this is the best case and should be used as a reference baseline. Results present in the following rows confirm that our *FfD* is able to generate high-quality faces, very similar to gray-level faces. Moreover, we note that the head pose network has the ability to generalize well on cross-dataset evaluations since we generally obtain a good accuracy even with different types of face images as input. The *Face-from-Depth* network has been created to this goal, even if the output is not always realistic and visually pleasant: however, the promising results confirm their positive contribution in the estimation of the head pose.

Head Pose Estimation. An ablation study of *POSEidon*$^+$ framework on *Pandora* is conducted and results are reported in Table 4.4, providing mean and standard deviation of the estimation errors obtained on each angle and for each system configuration. Similar to Fanelli *et al.* [64], we also report the mean accuracy as

percentage of good estimations (*i.e.*, angle error below 15°).
The first row of Table 4.4 shows the performance of a baseline system, obtained using the head pose estimation network only, and input depth frames are directly fed to the network without processing and cropping the input around the head. As expected, results are reasonable proving the ability of the deep network to extract useful features for head pose estimation from whole images.
The cropping step is included instead in the other rows, using the ground truth head position as the center and the cropping method described in Section 4.1.1. All three branches (i.e., depth, *FfD*, and Motion Images) of *POSEidon*$^+$ framework are individually evaluated. In particular, the fifth row includes an indirect evaluation of the reconstruction capabilities of the *Face-from-Depth* network. The same network trained and tested on the original gray level images performs similarly to the one trained and tested on *FfD* outputs (Row 3). The similar results confirm that the image reconstruction quality is sufficiently accurate, at least for the pose estimation task.
Results obtained using couples of networks are shown in rows 6 and 7, exploiting concatenation to merge the final layers of each component. Finally, the last row reports the performance of the complete framework. To merge layers, we use a *conv* fusion of couples of input types, followed by the *concat* step. We found that it is the best combination, as described in [20]. Even if the choice of the fusion method has a limited effect (as deeply investigated in [200, 68]), the most significant improvement of the system is reached by combining and exploiting the three input types together.

Figure 4.9 shows a comparison of the performance provided by each trident component: each graph plots the error distribution of a specific network with respect to the ground truth value. Depth data allows reaching the lowest error rates for frontal heads, while the other input data types are better in presence of rotated poses. The graphs highlight the averaging capabilities of *POSEidon*$^+$ too.
Furthermore, in Table 4.5 we compare best performance of *POSEidon*$^+$ on *Pandora* dataset, obtained exploiting the *FfD* network proposed in this work and the previous one described in [20]. We also evaluate *POSEidon*$^+$ replacing the central CNN (see Fig. 4.4) trained on reconstructed face images with the same CNN but trained on gray-level images (this experiment is here referred as *POSEidon**). Results confirm that the proposed *POSEidon*$^+$ overcomes our preliminary work. The overall quality of reconstructed face images is confirmed and also the feasibility to train and test the pose network on different dataset without a significant drop in performance.

Finally, we compare the results of *POSEidon*$^+$ with the state-of-art on the

Table 4.6: Estimation errors and mean accuracy of the shoulder pose estimation on *Pandora*

Parameters		**Shoulders**			**Accuracy**
R_x	R_y	Pitch	Roll	Yaw	
No crop		2.5 ± 2.3	3.0 ±2.6	3.7 ± 3.4	0.877
700	250	2.9 ± 2.6	2.6 ±2.5	4.0 ± 4.0	0.845
850	250	2.4 ± 2.2	2.5 ±2.2	3.1 ± 3.1	0.911
850	500	**2.2 ± 2.1**	**2.3 ±2.1**	**2.9 ± 2.9**	**0.924**

Biwi dataset. Due to the lack of a common validation and test protocol, Table 4.1 is split accordingly to the evaluation procedures adopted, in order to allow fair comparisons. For each validation procedure, we report results of *POSEidon*$^+$. In particular, we implement a 2-folds (half subjects in train and half in test), 4-folds, 5-folds (as adopted in the original works [66, 64], respectively) and 8-folds subject independent cross evaluations. We also conduct the *Leave-One-Out* (LOO) validation protocol. We dedicate the last section of Table 4.1 also for those methods that do not follow a standard evaluation procedure since they create a fixed or random [1] sets with a limited number of subjects (or sequences) to test their systems. Besides, we note that a fair comparison with methods reported in the top part of Table 4.1 is not possible since they exploit all sequences of *Biwi* dataset for test, while deep learning approaches need a certain amount of training data.
Results confirm the excellent performance of *POSEidon*$^+$ and the generalization ability across different training and testing subsets with different validation protocol. The system overcomes all the reported methods, included our previous proposal [20]. The average error is lower than other approaches, even those are not using all the frames available on *Biwi* dataset (some works exclude the frames on which the face detection fails [66, 64]).

Shoulder Pose Estimation. The network performing the shoulder pose estimation has been tested on *Pandora* only, due to the lack of the corresponding annotation in the other datasets. Results are reported in Table 4.6.
In particular, we conduct evaluation on different input types, varying the values R_x and R_y (cf. Section 4.1.3) that affect head and shoulder crops. We test also the shoulder pose network using the whole input depth frame, without any crop. The reported results are very promising, reaching an accuracy of over 92%.

Complete pipeline. In order to have a fair comparison, results reported in

Table 4.7: Results on *Biwi*, *ICT-3DHP* and *Pandora* dataset of the complete *POSEidon*$^+$ pipeline (*i.e.*, head localization, cropping and pose estimation).

Dataset	**Local.**	**Head**		
		Pitch	Roll	Yaw
Biwi	3.27±2.19	1.5±1.4	1.6±1.6	2.2±2.0
ICT-3DHP	-	4.9±4.2	3.5±3.4	6.8±6.0
Pandora	4.27±3.25	7.3±8.2	4.6±4.5	10.3±11.4

Tables 4.1 and 4.4 are obtained using the ground truth head position as input to the crop procedure. We finally test the whole pipeline, including the head localization network described in section 4.1.1, using also *ICT-3DHP* dataset. The mean error of the head localization (in pixels) and the pose estimation errors are summarized in Table 4.7. Sometimes, the estimated position generates a more effective crop of the head and, as a result, the whole pipeline performs better on the head pose estimation over the *Biwi* dataset. *POSEidon*$^+$ reaches valuable results also on the *ICT-3DHP* dataset and it provides comparable results with respect to state-of-the-art methods working on both depth and RGB data (4.9±5.3, 4.4±4.6, 5.1±5.4 [232], 7.06, 10.48, 6.90 [11], for pitch, roll and yaw respectively). We note that *ICT-3DHP* does not include the head center annotation, but the position of the device used to acquire pose data placed on the back of the head, and this partially compromises the performance of our method. Besides, we can not suppose a coherency between the annotations obtained with different IMU devices, in particular regarding the definition of the null position (*i.e.*, when the head angles are equal to zero).
The complete framework – except for the *FfD* module – has been implemented and tested on a desktop computer equipped with a *NVidia Quadro k2200* GPU board and on a laptop with a *NVidia GTX 860M*. Real-time performance has been obtained in both cases, with a processing rate of more than 30 frames per second. The whole system has been tested instead on a *Nvidia GTX 1080* and is able to run at more than 50 frames per second. Some examples of the system output are reported in Figure 4.1. In addition, the original depth map, the *Face-from-Depth* reconstruction and the motion data given in input to *POSEidon*$^+$ are placed on the left of each frame.

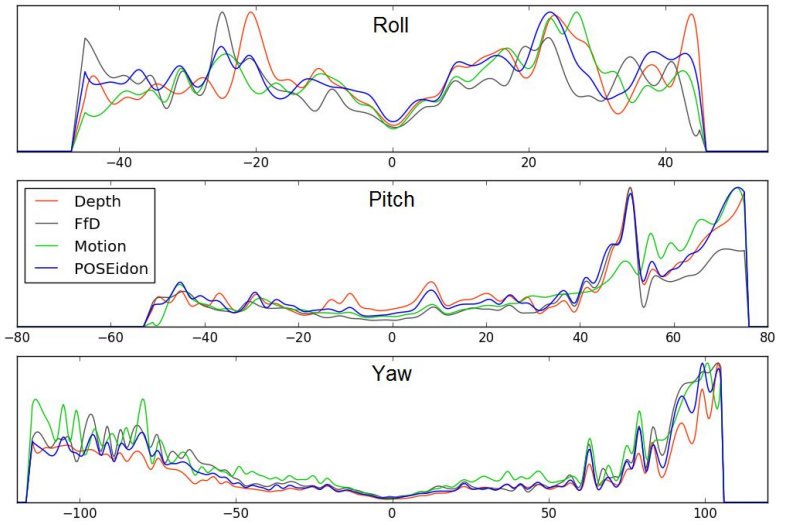

Figure 4.9: Error distribution of each *POSEidon*$^+$ components on *Pandora* dataset. On x-axis are reported the ground truth angles, on y-axis the distribution of error for each input type.

4.2 Attribute Recognition and Landmark Localization

As previously shown, Generative Adversarial Networks (GANs) have been adopted as a viable and efficient solution for the Image-to-Image translation task, or rather the ability to transform images into other images across domains, according to a specific training set. Initially, Autoencoders, and in particular Convolutional Autoencoders [167], have been investigated and designed for several image processing tasks, such as image restoration [162], deblurring [14], and for image transformations such as image inpainting [86] or image style transformation. They have been used also as transfer learning mechanism for the Domain Transfer task: for sensor to image transformation [246] or from depth to gray-level images of faces [20]. As mentioned above, the Goodfellow's proposal of GANs [84] became the reference architecture for unsupervised generative modeling and for sampling new images from the underlying distribution of an unlabeled dataset by exploiting the joint capabilities of a Generative and a Discriminative Networks

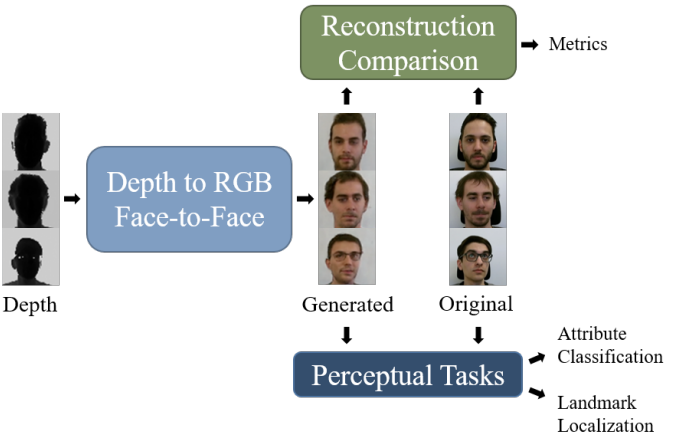

Figure 4.10: Overview of Reconstruction Comparison and Probe Perceptual Tasks for performance evaluation.

[210]. Furthermore, Conditional GANs [107, 61] provided conditional generative models by conditioning the sampling process with a partially observed input image. Several experiments show the power and effectiveness of conditional GANs, as for instance to improve resolution or to provide de-occlusion of images of people [61].

In this work, we explore the capability of face-to-face domain translation exploiting conditional GANs. The ability of a network to hallucinate and define a face aspect (in color or gray level), starting from a range map, could be a useful basic step for many computer vision and pattern recognition tasks, from biometric to expression recognition, from head pose estimation to interaction, especially in those contexts where intensity or color images cannot be recorded, for instance when shadows, light variations or darkness make the luminance and color acquisition not feasible enough. Our contribution is the definition of a *Conditional Generative Adversarial Network* that, starting from an annotated dataset with coupled depth and RGB faces (acquired by RGB-D sensors), learns to generate a plausible RGB face from solely the depth data. The network learns a proper transformation across the color and depth domains. Nevertheless, in generative settings, the result is likely a plausible face which could be qualitatively

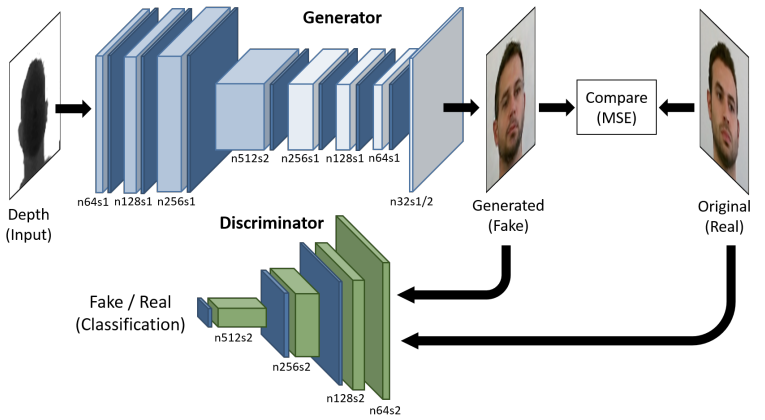

Figure 4.11: Training schedule for Conditional GANs. The Discriminator learns to classify between generated fake images and real images while the Generator learns to fool the Discriminator. For each layer, the image provide information about number of filters (n) and stride (s).

satisfactory (*e.g.*, the Discriminator network is fooled by it) but it is objectively difficult to properly measure the adherence to the conditioned input.
Therefore, another important contribution of our proposal is the adoption of some vision tasks as *Perceptual Probes* for performance evaluation, under the assumption that the domain translation task is viewed as an initial step of a more complex visual recognition task. We assess that the face-to-face translation is acceptable if the new generated RGB face (from depth input) exhibits similar proprieties of other RGB-native faces in the selected probe perceptual tasks (*i.e.*, categorical attributes are maintained across domains). In accordance with this assumption, we will provide several experiments to test the proposed solution: we will use two different perceptual probes – namely, a network for face attribute classification and a method for landmark extraction – and we will evaluate how these tasks perform on generated faces. The overview of our Probe Perceptual Task is depicted in Figure 4.10. Results are really encouraging so that this approach could be a first attempt to "see and recognize faces in the dark", in analogy to how blind people captures the appearance only by touching a face and sensing the depth shape.

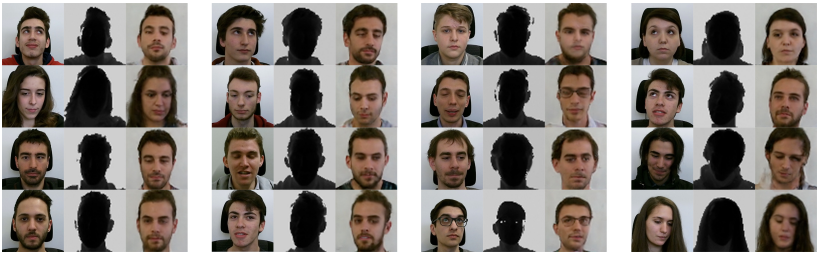

Figure 4.12: Best results on the *MotorMark* dataset. For each triplets of images: (Leftmost) the original image; (Middle) the input depth map; (Rightmost) the Generated face image.

4.2.1 Proposed Method

A general view of the proposed method is depicted in Figure 4.11. It consists of a GAN trained and tested on two different datasets, detailed in the following section. GANs are generative models that learn a mapping from random noise vector z to output image y: $G : z \rightarrow y$ [84]. Conditional GANs instead are generative models introduced by [107] that learn a mapping from an observed image x and random noise z to an output image y: $G : \{x, z\} \rightarrow y$. Like GANs, they are composed of two components: a Generator G and a Discriminator D. The Generator G is trained to generate outputs that are indistinguishable from "real" by the adversarially trained Discriminator D which is trained to recognize the Generator's "fake" images from the "real" ones.

Using random noise as input, the generator G creates completely new samples, drawn from a probability distribution that approximates the distribution of the training data. This procedure leads to a non-deterministic behavior, that is undesired for our goal. By removing the noise z, the probability distribution approximated by the model becomes a delta function with the property of preserving a deterministic behavior. Deterministic Conditional GANs (det-cGAN) thus learn a mapping from observed image x to output image y: $G : x \rightarrow y$.

Framework

The main goal is to train a generative function G capable of estimating the RGB face appearance I^{gen} from the corresponding depth input map I^{dpt} with the objective of reproducing the original image I^{rgb} associated with the depth map.

Table 4.8: Evaluation metrics computed on the generated RGB face images with *MotorMark* dataset. Starting from left are reported $L1$ and $L2$ distances, absolute and squared error differences, root-mean-squared error and finally the percentage of pixels under a defined threshold.

Method	**Norm ↓**		**Difference ↓**		**RMSE ↓**			**Threshold ↑**		
	L_1	L_2	Abs	Squared	linear	log	scale-inv	1.25	2.5	3.75
Autoencoder	39.80	6327	2.21	273.33	58.74	1.248	1.791	1.389	1.878	2.120
pix2pix [107]	37.77	6150	2.06	253.11	56.01	1.240	1.846	1.400	1.882	2.157
Our	**37.12**	**6021**	**2.05**	**245.88**	**54.86**	**1.222**	**1.749**	**1.423**	**1.914**	**2.188**
Our (Binary Maps)	43.58	6868	2.45	320.93	62.53	1.320	1.830	1.319	1.778	2.047

To this aim, we train a Generator Network as a feed-forward CNN G_{θ_g} with parameters θ_g. For N training pairs images (I^{dpt}, I^{rgb}) we solve:

$$\hat{\theta}_g = \arg\min_{\theta_g} \frac{1}{N} \sum_{n=1}^{N} Loss_G \left(G_{\theta_g}(I_n^{dpt}), I_n^{rgb} \right). \tag{4.8}$$

We obtained $\hat{\theta}_g$ by minimizing the loss function defined at the end of this subsection. Following the det-cGAN paradigm we further define a Discriminator Network D_{θ_d} with parameters θ_d that we train alongside G_{θ_g} with the aim of solving the adversarial min-max problem:

$$\min_{\theta_g} \max_{\theta_d} \mathbb{E}_{I^{rgb} \sim p_{data}(I^{rgb})} [\log D(I^{rgb})] + \mathbb{E}_{I^{gen} \sim p_{gen}(I^{gen})} [\log 1 - D(G(I^{dpt}))] \tag{4.9}$$

where $D(I^{rgb})$ is the probability of I^{rgb} being a "real" image while $1 - D(G(I^{dpt}))$ is the probability of $G(I^{dpt})$ being a "fake" image. The main idea behind this min-max formulation is that it gives the possibility to train a generative model G with the target of fooling the discriminator D, which is adversarially trained to distinguish between generated "fake" images and "real" ones. With this approach, we achieve a generative model capable of learning solutions that are highly similar to "real" images, thus indistinguishable from the Discriminator D.

As a possible drawback, those solutions could be highly realistic thanks to D but unrelated to the input. A generated output could be a realistic face image with very different visual attributes and different pose with respect to the original image. This setup does not guarantee, for example, that a depth map of a girl with wavy

hair looking to the right will generate an RGB image preserving those features. In order to tackle this problem, we mixed the Generator loss function $Loss_G$ with a more canonical loss such as MSE. Borrowing the idea from [61], we propose a Generator loss that is a weighted combination of two components:

$$Loss_G = \lambda Loss_{MSE} + Loss_{adv} \tag{4.10}$$

where $Loss_{MSE}$ is calculated using the mean squared errors of prediction (MSE) which measure the discrepancy between the generated image I^{gen} and the ground truth image I^{rgb} associated with the corresponding input depth map I^{dpt}. The MSE component is subject to a multiplication factor λ which controls its impact during training. The $Loss_{adv}$ component is the actual adversarial loss of the framework which encourages the Generator to produce perceptually good solutions that reside in the manifold of face images. The loss is defined as follows:

$$Loss_{adv} = \sum_{n=1}^{N} -\log(D(G(I^{dpt}))) \tag{4.11}$$

where $D(G(I^{dpt})$ is the probability of the Discriminator labeling the generated image $G(I^{dpt})$ as being a "real" image. Rather than training the Generator to minimize $\log(1 - D(G(I^{dpt})))$ we train G to minimize $\log(D(G(I^{dpt})))$. This objective provides strongest gradients early in training [84]. The combination of those two component grants the required behavior: the Generator has not only to fool the Discriminator but has to be near the ground truth output in an MSE sense.

Architecture

The task of Image-to-Image translation can be expressed as finding a mapping between two images. In particular, for the specific problem we are considering, the two images share the same underlying structure despite differing in surface appearance. Therefore, the structure in the input depth image is roughly aligned with the structure in the output RGB image. In fact, both images are representing the same subject in the same pose thus details like mouth, eyes, and nose share the same location through the two images. The generator architecture was designed following those considerations.

A recent solution [107] to this task adopted the "U-Net" [228] architecture with skip connections between mirrored layers in the encoder and decoder segments in order to shuttle low-level information between input and output directly across the network. We found this solution less profitable because the strictly underlying

structural coherence between input and output makes the network use the skip connections to jump at easier but not optimal solutions and ignoring the main network flow.

Consequently, our architecture implementation follows the *FfD* implementation in [19]. We relaxed the structure of the classical hourglass architecture performing less upsampling and downsampling operations in order to preserve the structural coherence between input and output. We found that using the half of feature maps described in [19] at each layer in both Generator and Discriminator networks sped up the training without a significant reduction of qualitative performance.

We propose the Generator's architecture depicted in Figure 4.11. Specifically, in the encoder, we used three convolutions followed by a strided convolution (with stride 2, in order to reduce the image resolution). The decoder uses three convolutions followed by a fractionally strided convolution (also known in literature as transposed convolutions) with stride $1/2$ to increase the resolution, and a final convolution. Leaky ReLU is adopted as activation function in the encoding stack while ReLU is preferred in the decoding stack. Batch normalization layers are adopted before each activation, except for the last convolutional layer which uses the Tanh activation. The number of filters follows a power of 2 pattern: from 64 to 512 in the encoder and from 256 to 32 in the decoder. All convolutions use a kernel of size 5×5. The Discriminator architecture is similar to the Generator's encoder in terms of number of filters and activations functions but uses only strided convolutional layers with stride 2 to halve the image resolution each time the number of filters is doubled. The resulting 512 feature maps are followed by one sigmoid activation to obtain a probability useful for the classification problem.

Training Details

We trained our det-cGAN with 64×64 resized depth maps as input and simultaneously providing the original RGB images associated with the depth data in order to compute the MSE loss. We adopted the standard approach in [84] to optimize the network alternating gradient descent updates between the generator and the discriminator with K = 1. We used mini-batch SGD applying the *Adam* solver with momentum parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$. In our experiments we chose a λ value of 10^{-1} in Equation (4.10) and a batch size of 64. Some best results are presented in Figure 4.12.

4.2.2 Experiments

Generally, evaluating the quality of reconstructed images is still an open problem, as reported in [107]. Traditional metrics such as $L1$ distance are not sufficient to assess joint statistic on the produced images, and therefore do not extrapolate the

Figure 4.13: Visual examples of generated images that preserve (left column) and do not preserve (right column) some attributes.

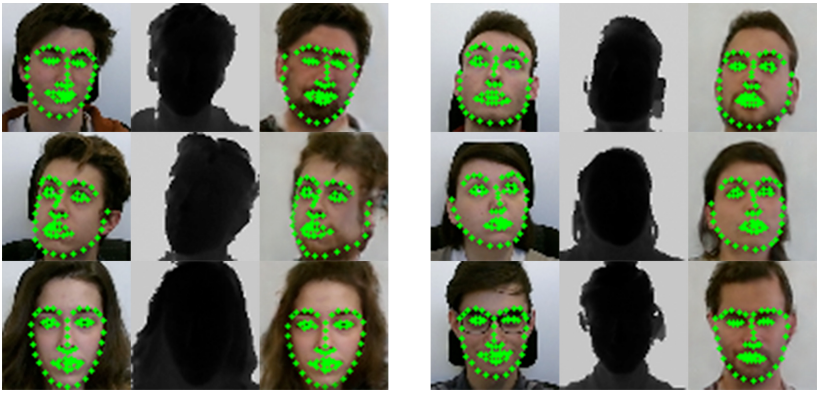

Figure 4.14: Visual examples of landmark predictions on real and generated images.

Table 4.9: Per attribute concordance between the true RGB face and the hallucinated one using VGG-Face CNN.

Attribute	Accuracy	Precision	Recall	F1
Male	90.30	95.51	93.49	94.49
Young	93.01	97.69	95.09	96.37
Mouth Open	82.86	92.16	51.07	65.71
Smiling	96.25	99.54	66.48	79.72
Wearing Hat	98.40	99.38	58.05	73.29
Wavy Hair	98.46	95.28	48.44	64.24
No Beard	48.18	63.89	40.88	49.86
Straight Hair	79.12	07.78	57.76	13.71
Eyeglasses	80.12	24.91	08.14	12.27

full structure of the result. In order to more holistically investigate the capabilities of our network to synthesize RGB face images directly from depth maps, a reconstruction comparison and two perceptual probes are performed. Firstly, we compared the performance of the proposed model with other *Image-to-Image* recent methods present in the literature, through metrics directly calculated over the reconstructed images. Secondly, we measured the capability of the proposed

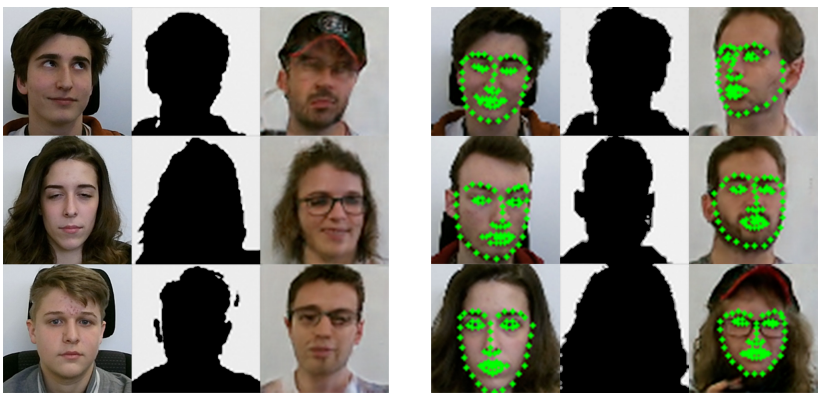

Figure 4.15: Visual examples of issues using binary maps instead of depth maps: attributes are not preserved (left column) and landmark localization is not precise (right column).

network of being able to preserve original facial attributes, like wearing hat and smiling, by exploiting a classification network trained with RGB face images. Thirdly, we measured whether or not reconstructed RGB face images are realistic enough that an off-the-shelf landmark localization system is able to localize accurate key-points. Eventually, in order to investigate how much the depth map information impacts the reconstruction task, we repeated the previous experiments testing our network trained with binary maps derived from the original depth maps.

Datasets

Experiments are conducted exploiting two publicly available datasets: *Pandora* [20] and *MotorMark* [72]. *Pandora* contains more than 250k frames, splitted into 110 annotated sequences of 22 different actors (10 males and 12 females), while *MotorMark* is composed of more than 30k frames of 35 different subjects, guaranteeing a great variety of face appearances. Subjects can wear garments and sunglasses and may perform driving activities actions like turning the steering wheel, adjust the rear mirror and so on. Both datasets have been acquired with a *Microsoft Kinect One*. In our experiments, *Pandora* has been used as the training set and *MotorMark* as the test set, performing a cross-dataset validation of the proposed method.

Reconstruction Comparison

Here, we check the capabilities of the proposed network to reconstruct RGB images from the correspondent depth ones. We exploited the metrics described in [57]: these metrics were originally used to evaluate depth images generated from RGB image sources. Results are reported in Table 4.8. In particular, we compared our generative method with two other techniques: an Autoencoder trained with the same architecture as our Generator network, and *pix2pix* [107], a recent work that exploits the Conditional GAN framework. In the last line of Table 4.8, is also reported the comparison with our network trained on binary maps, detailed at the end of this section. As shown, results confirm the superior accuracy of the presented method.

Attribute Classification

In the previous section, we checked the overall quality of the reconstructed RGB images. Here, we focus on the capability of our network to generate face images

Table 4.10: Quantitative comparison about the average attributes concordance between true and hallucinated RGB faces.

Method	Accuracy	Precision	Recall	F1
Autoencoder	75.21	61.84	40.55	51.38
pix2pix	84.57	74.42	56.01	60.78
Our	**85.19**	**75.13**	**57.71**	**61.07**
Our (Binary Maps)	60.51	49.48	29.12	42.76

that specifically preserve the facial attributes of the original person. To this end, we exploited a pre-trained network, the *VGG-Face* CNN [201], trained on RGB images for face recognition purposes. In order to extrapolate only the attributes that can be carried by depth information, we fine-tuned the network with the *CelebA* Dataset [154].

By observing Table 4.9, it is evident the good capability of the network to preserve gender, age, pose, and appearance attributes. Nevertheless, the depth sensor resolution fails at modeling hair categories such as curly or straight and glasses since such details are not always correctly captured in terms of depth. Glasses lenses, for example, are neglected by IR sensors and significantly captured only when the glasses structure is solid and visible. In all the other cases they tend to be confused by the network with the ocular cavities. Nonetheless, Table 4.10 exhibits the superiority of our proposal against state of the art generative networks also in attribute preservation. Moreover in Figure 4.13 are presented both successful and failure cases.

Landmark Localization

The intuition behind this experiment is that if the synthesized images are realistic and accurate enough, then a landmark localization method trained on real images will be able to localize key-points also on the generated images. To this aim, we exploited the algorithm included in the *dLib* libraries [123], which gives landmark positions on RGB images. In Figure 4.14 qualitative examples that highlight the coherence of landmark predictions between original and generated images are presented. The last column of Table 4.11 reports, for each method, the average $L2$ Norm between the position of landmarks predicted and the ground truth provided by the dataset. The results show that our method is able to produce outputs that can fool an algorithm trained on RGB face images.

Table 4.11: Quantitative comparative results of our proposal against the Autoencoder and pix2pix baselines in terms of face detector accuracy and landmark localization.

Method	**Accuracy**	$L2$ **Norm**
Autoencoder	54.03	2.219
pix2pix	85.21	2.201
Our	**86.86**	**2.089**
Our (Binary Maps)	62.37	2.980

Binary Maps

An ablation study is conducted to investigate the importance of depth information, by training our network providing as input binary maps instead of depth maps. Binary maps were gathered thresholding the depth maps. Figure 4.15 shows examples where the reconstructed face images are not coherent with the original images in terms of attributes preservation and landmark position. At the end of Tables 4.8, 4.10 and 4.11 are reported the results of the previous experiment where we used binary maps instead of depth maps. Results show that the depth information has a fundamental importance in the face generation task, to preserve coherent facial attributes and head pose orientation.

Chapter 5

Conclusions

The aim of this thesis - and of the research work done during my PhD - was to enable a deeper understanding of human behaviour in surveillance scenarios. This aim disclosed a variety of open problems in computer vision that will need to be addressed before reaching any satisfactory answer. In the quest for reaching such objective, I contributed to five of those open problems, namely 2D pose estimation and tracking, 3D pose estimation, attribute recognition, 3D people detection and head pose estimation. The synthetically generated data collected during my PhD played a pivotal role by enabling new possibilities that go beyond the scope of this thesis.

2D Pose Estimation and Tracking

In this thesis, 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 integrated temporal coherency and proposed 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.

3D Pose Estimation

In this work we presented a single-shot bottom-up approach for multi-person 3D HPE suitable for both crowded surveillance scenarios and for simpler, even single person, contexts without any changes. Our LoCO approach allowrd us to exploit volumetric heatmaps as a ground truth representation for the 3D HPE 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 showed state-of-the-art performance on all the considered datasets. We also believe that this new simple compression strategy can foster future research by enabling the full potential of the volumetric heatmap representation in contexts where it was previously intractable.

Attributes Recognition

As for attribute recognition, we presented the use of GANs for image enhancing in people attributes classification. Our generator network has been designed to overcome a common problem in surveillance scenarios, namely people occlusion. Experiments showed that jointly enhancing images before feeding them to an attribute classification network can improve the results even when input images are affected by this issue. We think that this line of work can foster research about the problem of attribute classification in surveillance contexts, where camera resolution and positioning cannot be neglected.

3D People Detection

For 3D people detection we proposed a simple and effective system that deals with the COVID-19 emergency by providing a social distancing tool that can prevent the spread of the infection. We validated it using a highly challenging benchmark, obtaining a lower bound on the performance of the method. We believe that our

system can be a practical solution to an important problem, hoping to see areas less crowded than the JTA dataset in the near future.

Head Pose Estimation

An end-to-end framework to monitor the driver's body pose called *POSEidon*$^+$ has been presented. In particular, a new *Face-from-Depth* architecture has been proposed, based on a Deterministic Conditional GAN approach, to convert depth faces in gray-level images and supporting head pose prediction.
The system is based only on depth images, no previous computation of specific facial features is required and has shown real-time and impressive results with two public datasets. All these aspects make the proposed framework suitable to particular challenging contexts, such as automotive. Since the system has been developed with a modular architecture, each module can be used as single or in combination, reaching worst but still satisfactory performances. This work provides a comprehensive review and a comparison of recent state-of-art works and can be used as a brief review to understanding the current state of the 3D head pose estimation task.

Acknowledgements

This thesis has been made possible by the help and contributions of many people. We do our best to acknowledge them all here.

The author would like to acknowledge my supervisor Prof. Rita Cucchiara for constantly pushing me to my best, inspiring and motivating me over the last three years.

The author would like to acknowledge my co-supervisor Simone, precious ally during both hard and good times.

The author would like to acknowledge Fabio who has been the best you could ask for as a colleague and friend. I am pretty sure that, without him, none of what is written in this thesis would have been possible.

The author would like to acknowledge Riccardo because he was always there

when I needed him. Thank you for being such a generous guy.

The author would like to acknowledge Stefano for hosting me at his home during my internship and teaching me how to survive in a foreign country.

The author would like to acknowledge Guido for the fruitful collaborations related to the Face-from-Depth line of research.

The author would like to acknowledge Gianluca and Federico for their tireless work, especially on modding and on paper experiments.

The author would like to acknowledge every colleague from the AImageLab for making me feel at home since the day I walked in.

The author would like to acknowledge the Italian MIUR, Ministry of Education, Universities and Research and Regione Emilia Romagna for their financial support.

The author would like to acknowledge Panasonic for the wonderful opportunities resulting from my internship.

The author would like to acknowledge Rockstar Games for creating the videogame Grand Theft Auto V, fundamental for the data collection carried out during my PhD period.

Appendix A

List of publications

In this section we briefly report the research papers published during the PhD period (including preprint if proceeding not available).

- Compressed Volumetric Heatmaps for Multi-Person 3D Pose Estimation
  Matteo Fabbri, Fabio Lanzi, Simone Calderara, Stefano Alletto, Rita Cucchiara
  Computer Vision and Pattern Recognition. 2020.
- Inter-Homines: Distance-Based Risk Estimation for Human Safety
  Matteo Fabbri, Fabio Lanzi, Riccardo Gasparini, Simone Calderara, Lorenzo Baraldi, Rita Cucchiara
  arXiv preprint arXiv:2007.10243. 2020.
- Can Adversarial Networks Hallucinate Occluded People With a Plausible Aspect?
  Federico Fulgeri, Matteo Fabbri, Stefano Alletto, Simone Calderara, Rita Cucchiara
  Computer Vision and Image Understanding. 2019.
- Face-from-Depth for Head Pose Estimation on Depth Images
  Guido Borghi, Matteo Fabbri, Roberto Vezzani, Simone Calderara, Rita Cucchiara
  IEEE Transactions on Pattern Analysis and Machine Intelligence. 2018.

- 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
  IEEE European Conference on Computer Vision. 2018.

- Domain Translation with Conditional GANs: from Depth to RGB Face-to-Face
  Matteo Fabbri, Guido Borghi, Fabio Lanzi, Roberto Vezzani, Simone Calderara, Rita Cucchiara
  ICPR International Conference on Pattern Recognition. 2018.

- Generative Adversarial Models for People Attribute Recognition in Surveillance
  Matteo Fabbri, Simone Calderara, Rita Cucchiara
  IEEE International Conference on Advanced Video and Signal based Surveillance. 2017.

Appendix B

Activities carried out during the PhD

Here we report research activities carried out during the 3 years of PhD.

Foreign collaborations

- Research Internship at Panasonic R&D Company of America. Mountain View - California (US), January - December 2019
- Research collaboration with Technical University of Munich (Prof. Laura Leal-Taixé) and TU Darmstadt (Prof. Stefan Roth), January 2020 - March 2021

Conferences, courses, seminars attended

Conferences and Tutorials

- International Conference on Computer Vision and Pattern Recognition (CVPR), virtual, 2020
- Panasonic Technological Symposium (PTS), Osaka, 2019

- International Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, 2019
- Barkeley Spring 2019 BAIR/BDD Retreat, Santa Rosa, 2019
- European Conference on Computer Vision (ECCV), Munich, 2018
- International Computer Vision Summer School (ICVSS), Sicily, 2018
- International Conference on Computer Vision (ICCV), Venice, 2017

Courses and seminars

- Managing the Company of the Future - Prof. Julian Birkinshaw - September 24th 2020
- Deep Learning for Fault Prediction - Prof. Roberto Paredes Palacios - February 2018
- Algoritmi Avanzati - Dr. Mauro Leoncini - September 2017
- Learn how to activate learn: Meta-learning approaches to (deep) active learning - Prof. Massimiliano Ruocco - May 9th, 2018
- Computational and experimental neuroscience toward artificial intelligence - Prof. Jonathan Mapelli - April 10th, 2018
- Deep learning technologies: from hardware components to vertical frameworks - Dr. Piero Altoè, NVIDIA - November 29th, 2017.

Award and Prizes

- Winner of Best Demo Award, ECCV 2020
- Winner of Best Paper Award, PTS 2019

Bibliography

[1] Byungtae Ahn, Jaesik Park, and In So Kweon. Real-time head orientation from a monocular camera using deep neural network. pages 82–96, 2014. 12, 90, 97, 109

[2] Giuseppe Amato, Luca Ciampi, Fabrizio Falchi, Claudio Gennaro, and Nicola Messina. Learning pedestrian detection from virtual worlds. In *Int. Conf. on Image Analysis and Proc.*, 2019. 7

[3] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. *CVPR*, 2014. 20

[4] Mykhaylo Andriluka, Umar Iqbal, Anton Milan, Eldar Insafutdinov, Leonid Pishchulin, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5167–5176, 2018. 18, 19, 20, 32

[5] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. People-tracking-by-detection and people-detection-by-tracking. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 9

[6] H. Arai, Y. Chayama, H. Iyatomi, and K. Oishi. 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)*, pages 5162–5165, July 2018. 36

[7] Sima Asadi, Nicole Bouvier, Anthony S Wexler, and William D Ristenpart. The coronavirus pandemic and aerosols: Does covid-19 transmit via expiratory particles?, 2020. 55

[8] S. H. Bae and K. J. Yoon. 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):595–610, March 2018. 9, 33

[9] Andrew D. Bagdanov, Iacopo Masi, and Alberto Del Bimbo. The florence 2d/3d hybrid face datset. In *Proc. of ACM Multimedia Int.'l Workshop on Multimedia access to 3D Human Objects (MA3HO'11)*. ACM Press, December 2011. 102

[10] Simon Baker, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *International Journal of Computer Vision*, 2011. 6

[11] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 3d constrained local model for rigid and non-rigid facial tracking. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 2610–2617, 2012. 11, 14, 97, 103, 110

[12] Tobias Bär, Jan Felix Reuter, and J Marius Zöllner. Driver head pose and gaze estimation based on multi-template icp 3-d point cloud alignment. In *2012 15th International IEEE Conference on Intelligent Transportation Systems*, pages 1797–1802, 2012. 13

[13] Luis Miguel Bergasa, Jesús Nuevo, Miguel A Sotelo, Rafael Barea, and María Elena Lopez. Real-time system for monitoring driver vigilance. *IEEE Transactions on Intelligent Transportation Systems*, 7(1):63–77, 2006. 90

[14] Siavash Arjomand Bigdeli and Matthias Zwicker. Image restoration using autoencoding priors. *arXiv preprint arXiv:1703.09964*, 2017. 111

[15] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proc. of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999. 12

[16] Amit Bleiweiss and Michael Werman. Robust head pose estimation by fusing time-of-flight depth and color. In *Proc. of IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pages 116–121, 2010. 14

[17] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, 2017. 8

[18] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision (ECCV)*, 2016. 10

[19] Guido Borghi, Matteo Fabbri, Roberto Vezzani, Simone Calderara, and Rita Cucchiara. Face-from-depth for head pose estimation on depth images. *arXiv preprint arXiv:1712.05277*, 2017. 117

[20] Guido Borghi, Marco Venturelli, Roberto Vezzani, and Rita Cucchiara. Poseidon: Face-from-depth for driver pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 13, 16, 90, 92, 97, 99, 105, 106, 107, 108, 109, 111, 120

[21] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2017. 7

[22] Guillem Brasó and Laura Leal-Taixé. Learning a neural solver for multiple object tracking. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2020. 6

[23] Michael D Breitenstein, Daniel Kuettel, Thibaut Weise, Luc Van Gool, and Hanspeter Pfister. Real-time face pose estimation from single range images. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 13

[24] Adrian Bulat and Georgios Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *European Conference on Computer Vision*, pages 717–732. Springer, 2016. 8

[25] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *European Conference on Computer Vision*, 2012. 6

[26] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2020. 5

[27] Qin Cai, David Gallup, Cha Zhang, and Zhengyou Zhang. 3d deformable face tracking with a commodity depth camera. In *European Conference on Computer Vision*, pages 229–242, 2010. 11, 14

[28] Chen Cao, Yanlin Weng, Stephen Lin, and Kun Zhou. 3d shape regression for real-time facial animation. *ACM Transactions on Graphics (TOG)*, 32(4):41, 2013. 12

[29] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, volume 1, page 7, 2017. 8, 9, 11, 18, 21, 24, 26, 30, 31, 36, 41

[30] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4733–4742, 2016. 8

[31] Florian Chabot, Mohamed Chaouch, Jaonary Rabarisoa, Céline Teuliere, and Thierry Chateau. Deep manta: A coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 8

[32] James Charles, Tomas Pfister, Derek Magee, David Hogg, and Andrew Zisserman. Personalizing human video pose estimation. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 3063–3072. IEEE, 2016. 18, 19

[33] Chao-Yeh Chen and Kristen Grauman. Inferring unseen views of people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2003–2010, 2014. 16

[34] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation= 2d pose estimation+ matching. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 10

[35] Jiawei Chen, Jonathan Wu, Kristi Richter, Janusz Konrad, and Prakash Ishwar. Estimating head pose orientation using extremely low resolution images. In *2016 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)*, pages 65–68, 2016. 13

[36] Qiang Chen, Junshi Huang, Rogerio Feris, Lisa M Brown, Jian Dong, and Shuicheng Yan. Deep domain adaptation for describing people based on fine-grained clothing attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5315–5324, 2015. 14

[37] Siyuan Chen, Francois Bremond, Hung Nguyen, and Hugues Thomas. Exploring depth information for head detection with depth images. In *Advanced Video and Signal Based Surveillance (AVSS), 2016 13th IEEE International Conference on*, pages 228–234. IEEE, 2016. 12

[38] Yi-Cheng Chen, Ping-En Lu, Cheng-Shang Chang, and Tzu-Hsuan Liu. A time-dependent sir model for covid-19 with undetectable infected persons. *arXiv:2003.00122*, 2020. 55

[39] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 41

[40] Q. Chu, W. Ouyang, H. Li, X. Wang, B. Liu, and N. Yu. Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4846–4855, Oct 2017. 9, 33

[41] D. Coppi, S. Calderara, and R. Cucchiara. Transductive people tracking in unconstrained surveillance. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(4):762–775, 2016. 71

[42] Błażej Czupryński and Adam Strupczewski. High accuracy head pose tracking survey. In *International Conference on Active Media Technology*, pages 407–420. Springer, 2014. 90

[43] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. Learning 3d human pose from structure and motion. In *European Conference on Computer Vision (ECCV)*, 2018. 10, 49

[44] Angela Dai and Matthias Nießner. 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In *European Conference on Computer Vision (ECCV)*, 2018. 37

[45] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005. 14

[46] A. Dehghan, Y. Tian, P. H. S. Torr, and M. Shah. Target identity-aware network flow for online multiple target tracking. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1146–1154, June 2015. 18

[47] Patrick Dendorfer, Aljoša Ošep, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, and Stefan Roth Laura Leal-Taixé. Motchallenge: A benchmark for single-camera multiple target tracking. *International Journal of Computer Vision (IJCV)*, 2020. 6

[48] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020. 6

[49] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 50, 77

[50] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, 2014. 75, 81

[51] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning to recognize pedestrian attribute. *arXiv preprint arXiv:1501.00901*, 2015. 14

[52] Yue Deng, Qionghai Dai, and Zengke Zhang. Graph laplace for occluded face completion and recognition. *IEEE Transactions on Image Processing*, 20(8):2329–2338, 2011. 15

[53] Yue Deng, Dong Li, Xudong Xie, Kin-Man Lam, and Qionghai Dai. Partially occluded face completion and recognition. In *IEEE International Conference on Image Processing*, pages 4145–4148. IEEE, 2009. 15

[54] Anup Doshi and Mohan M Trivedi. Head and eye gaze dynamics during visual attention shifts in complex environments. *Journal of vision*, 12(2):9–9, 2012. 90

[55] Vincent Drouard, Sileye Ba, Georgios Evangelidis, Antoine Deleforge, and Radu Horaud. Head pose estimation via probabilistic high-dimensional regression. In *IEEE International Conference on Image Processing*, pages 4624–4628, 2015. 13, 90, 97

[56] Vincent Drouard, Sileye Ba, and Radu Horaud. Switching linear inverse-regression model for tracking head pose. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pages 1232–1240. IEEE, 2017. 12, 97

[57] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*. 120

[58] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2366–2374. Curran Associates, Inc., 2014. 106

[59] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 2015. 5

[60] Matteo Fabbri, Simone Calderara, and Rita Cucchiara. Generative adversarial models for people attribute recognition in surveillance. In *Advanced Video and Signal Based Surveillance (AVSS), IEEE International Conference on*. IEEE, 2017. 71, 75, 78, 81, 82, 83, 86, 88

[61] Matteo Fabbri, Simone Calderara, and Rita Cucchiara. Generative adversarial models for people attribute recognition in surveillance. In *14th IEEE International Conference on Advanced Video and Signal based Surveillance*, 2017. 99, 105, 106, 112, 116

[62] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Stefano Alletto, and Rita Cucchiara. Compressed volumetric heatmaps for multi-person 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 56

[63] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Andrea Palazzi, Roberto Vezzani, and Rita Cucchiara. Learning to detect and track visible and occluded body joints in a virtual world. In *European Conference on Computer Vision (ECCV)*, 2018. 7, 8, 37, 44, 47, 56, 62, 67, 73

[64] Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc Van Gool. Random forests for real time 3d face analysis. *International Journal of Computer Vision*, 101(3):437–458, 2013. 13, 97, 107, 109

[65] Gabriele Fanelli, Juergen Gall, and Luc Van Gool. Real time head pose estimation with random regression forests. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 617–624, 2011. 12, 13, 102

[66] Gabriele Fanelli, Thibaut Weise, Juergen Gall, and Luc Van Gool. Real time head pose estimation from consumer depth cameras. In *Joint Pattern Recognition Symposium*, pages 101–110, 2011. 13, 97, 109

[67] Gunnar Farneback. Very high accuracy velocity estimation using orientation tensors, parametric motion, and simultaneous segmentation of the motion field. In *IEEE International Conference on Computer Vision*, volume 1, pages 171–177. IEEE, 2001. 100

[68] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. *arXiv preprint arXiv:1604.06573*, 2016. 108

[69] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 3057–3065, 2017. 18

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

[71] Mihai Fieraru, Anna Khoreva, Leonid Pishchulin, and Bernt Schiele. Learning to refine human pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018. 41

[72] Elia Frigieri, Guido Borghi, Roberto Vezzani, and Rita Cucchiara. Fast and accurate facial landmark localization in depth images for in-car applications. In *International Conference on Image Analysis and Processing*, 2017. 120

[73] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2016. 7

[74] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2012. 5

[75] Reza Shoja Ghiass, Ognjen Arandjelović, and Denis Laurendeau. Highly accurate and fully automatic head pose estimation from a low quality consumer-level rgb-d sensor. In *Proc. of the 2nd Workshop on Computational Models of Social Interactions: Human-Computer-Media Communication*, pages 25–34, 2015. 11, 14

[76] Amir Ghodrati, Xu Jia, Marco Pedersoli, and Tinne Tuytelaars. Towards automatic image editing: Learning to see another you. *arXiv preprint arXiv:1511.08446*, 2015. 16

[77] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2015. 7

[78] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014. 7, 8

[79] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Actions and attributes from wholes and parts. In *IEEE International Conference on Computer Vision*, pages 2470–2478, 2015. 15

[80] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Contextual action recognition with r* cnn. In *IEEE International Conference on Computer Vision*, pages 1080–1088, 2015. 15

[81] Georgia Gkioxari, Alexander Toshev, and Navdeep Jaitly. Chained predictions using convolutional neural networks. In *European Conference on Computer Vision*, pages 728–743. Springer, 2016. 9

[82] Thomas Golda, Tobias Kalb, Arne Schumann, and Jürgen Beyerer. Human pose estimation for real-world crowded scenarios. In *IEEE Int. Conf. Adv. Video and Signal Based Surv.*, 2019. 7

[83] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. 15, 71, 75, 76, 80

[84] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 98, 100, 111, 114, 116, 117

[85] Nicolas Gourier, Daniela Hall, and James L Crowley. Estimating face orientation from robust detection of salient facial structures. In *FG Net Workshop on Visual Observation of Deictic Gestures*, volume 6, 2004. 102

[86] Christine Guillemot and Olivier Le Meur. Image inpainting: Overview and recent advances. *IEEE signal processing magazine*, 31(1), 2014. 111

[87] Seyed Hamid Rezatofighi, Anton Milan, Zhen Zhang, Qinfeng Shi, Anthony Dick, and Ian Reid. Joint probabilistic data association revisited. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 18

[88] Emily M Hand and Rama Chellappa. Attributes for improved attributes: A multi-task network for attribute classification. *arXiv preprint arXiv:1604.07360*, 2016. 14

[89] Ankur Handa, Richard A Newcombe, Adrien Angeli, and Andrew J Davison. Real-time camera tracking: When is high frame-rate best? In *European Conference on Computer Vision*, 2012. 6

[90] Ankur Handa, Viorica Patraucean, Vijay Badrinarayanan, Simon Stent, and Roberto Cipolla. Understanding real world indoor scenes with synthetic data. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2016. 7

[91] Ankur Handa, Thomas Whelan, John McDonald, and Andrew J Davison. A benchmark for rgb-d visual odometry, 3d reconstruction and slam. 2014. 6

[92] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 62

[93] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017. 8, 73, 81

[94] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 8, 62

[95] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society, 2016. 77, 83

[96] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998. 14

[97] Samitha Herath, Mehrtash Harandi, and Fatih Porikli. Going deeper into action recognition: A survey. *Image and Vision Computing*, 60:4 – 21, 2017. Regularization Techniques for High-Dimensional Data Analysis. 71

[98] Mir Rayat Imtiaz Hossain and James J Little. Exploiting temporal information for 3d human pose estimation. In *European Conference on Computer Vision (ECCV)*, 2018. 10

[99] Peiyun Hu and Deva Ramanan. Bottom-up and top-down reasoning with hierarchical rectified gaussians. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5600–5609, 2016. 8

[100] Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang, and Alexander G Schwing. Sail-vos: Semantic amodal instance level video object segmentation-a synthetic dataset and baselines. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2019. 7

[101] Rui Huang, Shu Zhang, Tianyu Li, Ran He, et al. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. *arXiv preprint arXiv:1704.04086*, 2017. 16

[102] Eldar Insafutdinov, Mykhaylo Andriluka, Leonid Pishchulin, Siyu Tang, Evgeny Levinkov, Bjoern Andres, and Bernt Schiele. Arttrack: Articulated multi-person tracking in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 4327, 2017. 9

[103] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*, pages 34–50. Springer, 2016. 8, 9

[104] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014. 36, 37, 44

[105] Umar Iqbal and Juergen Gall. Multi-person pose estimation with local joint-to-person associations. In *European Conference on Computer Vision*, pages 627–642. Springer, 2016. 8, 9

[106] Umar Iqbal, Anton Milan, and Juergen Gall. Posetrack: Joint multi-person pose estimation and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, 2017. 9, 30

[107] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016. 15, 92, 96, 99, 105, 106, 112, 114, 115, 116, 118, 120

[108] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 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*, pages 5967–5976, 2017. 16, 74, 75, 78, 82, 86, 88

[109] Hamid Izadinia, Imran Saleemi, Wenhui Li, and Mubarak Shah. 2t: Multiple people multiple parts tracker. In *European Conference on Computer Vision*, pages 100–114. Springer, 2012. 9

[110] Arjun Jain, Jonathan Tompson, Yann LeCun, and Christoph Bregler. Modeep: A deep learning framework using motion features for human pose estimation. In *Asian conference on computer vision*, pages 302–315. Springer, 2014. 9

[111] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3192–3199. IEEE, 2013. 18, 19

[112] Qiang Ji, Zhiwei Zhu, and Peilin Lan. Real-time nonintrusive monitoring and prediction of driver fatigue. *IEEE transactions on vehicular technology*, 53(4):1052–1068, 2004. 90

[113] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 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*, pages 694–711, 2016. 76, 77

[114] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015. 37, 44

[115] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 44

[116] Felix Juefei-Xu, Eshan Verma, Parag Goel, Anisha Cherodian, and Marios Savvides. Deepgender: Occlusion and low resolution robust facial gender classification via progressively trained convolutional neural networks with attention. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 68–77, 2016. 14

[117] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 10

[118] Biliana Kaneva, Antonio Torralba, and William T Freeman. Evaluation of image features using a photorealistic virtual world. In *IEEE International Conference on Computer Vision*, 2011. 6

[119] Sertan Kaymak and Ioannis Patras. Exploiting depth and intensity information for head pose estimation with random forests and tensor models. In

Asian Conference on Computer Vision, pages 160–170. Springer, 2012. 14, 97

[120] James Kennedy. Particle swarm optimization. In *Encyclopedia of machine learning*, pages 760–766. Springer, 2011. 13

[121] William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *The royal society of london. Series A, Containing papers of a mathematical and physical character*, 1927. 57

[122] Khalil Khan, Massimo Mauro, Pierangelo Migliorati, and Riccardo Leonardi. Head pose estimation through multi-class face segmentation. In *Multimedia and Expo (ICME), 2017 IEEE International Conference on*, pages 175–180. IEEE, 2017. 13

[123] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009. 121

[124] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 100

[125] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 15

[126] Farid Abedan Kondori, Shahrouz Yousefi, Haibo Li, Samuel Sonning, and Sabina Sonning. 3d head pose estimation using the kinect. In *Proc. of International Conference on Wireless Communications and Signal Processing (WCSP)*, pages 1–4, 2011. 13

[127] Philipp Krähenbühl. Free supervision from video games. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2018. 6, 7

[128] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 94, 101, 102

[129] Abhijit Kundu, Yin Li, and James M Rehg. 3d-rcnn: Instance-level 3d object reconstruction via render-and-compare. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 8

[130] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiri Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. *CoRR*, abs/1711.07064, 2017. 76

[131] Brian Lao and Karthik Jagadeesh. Convolutional neural networks for fashion classification and object detection. 14

[132] Christoph Lassner, Gerard Pons-Moll, and Peter V Gehler. A generative model of people in clothing. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 6, 2017. 16

[133] Stéphane Lathuilière, Rémi Juge, Pablo Mesejo, Rafael Munoz-Salinas, and Radu Horaud. Deep mixture of linear inverse regressions applied to head-pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 13, 97

[134] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision*, 2018. 8

[135] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 105–114, 2017. 76

[136] Kyoungoh Lee, Inwoong Lee, and Sanghoon Lee. Propagating lstm: 3d pose estimation based on joint interdependency. In *European Conference on Computer Vision (ECCV)*, 2018. 10

[137] Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–42, 2015. 14

[138] Evgeny Levinkov, Jonas Uhrig, Siyu Tang, Mohamed Omran, Eldar Insafutdinov, Alexander Kirillov, Carsten Rother, Thomas Brox, Bernt Schiele, and Bjoern Andres. Joint graph decomposition & node labeling: Problem, algorithms, applications. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 8

[139] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *Asian Conference on Pattern Recognition*, pages 111–115. IEEE, 2015. 15

[140] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 111–115, 2015. 75

[141] Dangwei Li, Zhang Zhang, Xiaotang Chen, Haibin Ling, and Kaiqi Huang. A richly annotated dataset for pedestrian attribute recognition. *preprint arXiv:1603.07054*, 2016. 15, 75, 80, 81, 83

[142] Songnan Li, King Ngi Ngan, Raveendran Paramesran, and Lu Sheng. Real-time head pose tracking with online face template reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1922–1928, 2016. 14, 97

[143] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2014. 6

[144] Yining Li, Chen Huang, Chen Change Loy, and Xiaoou Tang. Human attribute recognition by deep hierarchical contexts. In *IEEE International Conference on Computer Vision*, pages 684–700. Springer, 2016. 15

[145] Yiming Liang and Yue Zhou. Multi-camera tracking exploiting person re-id technique. In Derong Liu, Shengli Xie, Yuanqing Li, Dongbin Zhao, and El-Sayed M. El-Alfy, editors, *Neural Information Processing*, pages 397–404, Cham, 2017. Springer International Publishing. 71

[146] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002. 13

[147] Mude Lin, Liang Lin, Xiaodan Liang, Keze Wang, and Hui Cheng. Recurrent 3d pose sequence machines. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 10

[148] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2017. 7

[149] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5, 21, 60, 81

[150] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *Advances in neural information processing systems*, pages 469–477, 2016. 16

[151] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, 2016. 7

[152] Xiabing Liu, Wei Liang, Yumeng Wang, Shuyang Li, and Mingtao Pei. 3d head pose estimation with convolutional neural network trained on synthetic images. In *IEEE International Conference on Image Processing*, pages 1289–1293, 2016. 13, 97

[153] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE international conference on computer vision*, pages 1–9, 2017. 75, 81, 83

[154] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 121

[155] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 16, 106

[156] Kok-Lim Low. Linear least-squares optimization for point-to-plane icp surface registration. *Techrep - Chapel Hill, University of North Carolina*, 4, 2004. 13

[157] Wenjie Luo, Bin Yang, and Raquel Urtasun. Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3569–3577, 2018. 40

[158] Diogo C Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 10

[159] Xiaolong Ma, Xiatian Zhu, Shaogang Gong, Xudong Xie, Jianming Hu, Kin-Man Lam, and Yisheng Zhong. Person re-identification by unsupervised video matching. *Pattern Recognition*, 65:197 – 210, 2017. 71

[160] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015. 77

[161] Sotiris Malassiotis and Michael G Strintzis. Robust real-time 3d head pose estimation from range data. *Pattern Recognition*, 38(8):1153–1165, 2005. 13, 90

[162] Xiao-Jiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using convolutional auto-encoders with symmetric skip connections. *arXiv preprint arXiv:1606.08921*, 2016. 111

[163] Javier Marin, David Vázquez, David Gerónimo, and Antonio M López. Learning appearance in virtual scenarios for pedestrian detection. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2010. 7

[164] Manuel Martin, Florian van de Camp, and Rainer Stiefelhagen. Real time head model creation and head pose estimation on consumer depth cameras. In *Proceedings of the 2014 2Nd International Conference on 3D Vision - Volume 01*, 3DV '14, pages 641–648, Washington, DC, USA, 2014. IEEE Computer Society. 97

[165] Sujitha Martin, Kevan Yuen, and Mohan M Trivedi. Vision for intelligent vehicles & applications (viva): Face detection and head pose challenge. In *Intelligent Vehicles Symposium (IV), 2016 IEEE*, pages 1010–1014. IEEE, 2016. 102

[166] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *International Conference on Computer Vision (ICCV)*, 2017. 10, 43, 46

[167] Jonathan Masci, Ueli Meier, Dan Cireşan, and Jürgen Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In *International Conference on Artificial Neural Networks*, pages 52–59. Springer, 2011. 16, 106, 111

[168] Yoshio Matsumoto and Alexander Zelinsky. An algorithm for real-time stereo vision implementation of head pose and gaze direction measurement. In *Proc. of IEEE International Conference on Automatic Face and Gesture Recognition*, pages 499–504, 2000. 12

[169] Nikolaus Mayer, Eddy Ilg, Philipp Fischer, Caner Hazirbas, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. What makes good synthetic training data for learning disparity and optical flow estimation? *International Journal of Computer Vision*, 2018. 6

[170] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2016. 7

[171] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *International Conference on 3D Vision (3DV)*, 2017. 10

[172] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *International Conference on 3D Vision (3DV)*, 2018. 11, 43, 45, 49

[173] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 2017. 10, 43, 45, 49

[174] Gregory P Meyer, Shalini Gupta, Iuri Frosio, Dikpal Reddy, and Jan Kautz. Robust model-based 3d head pose estimation. In *IEEE International Conference on Computer Vision*, pages 3649–3657, 2015. 11, 13, 90, 97

[175] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. MOT16: A benchmark for multi-object tracking. *arXiv: 1603.00831*, 2016. 18, 32

[176] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 6, 19

[177] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 16, 74

[178] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10133–10142, 2019. 10, 49

[179] Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 10

[180] Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Jana Kosecka. 3d bounding box estimation using deep learning and geometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 8

[181] Sankha S Mukherjee and Neil Martin Robertson. Deep head pose: Gaze-direction estimation in multimodal video. *IEEE Transactions on Multimedia*, 17(11):2094–2107, 2015. 14

[182] Erik Murphy-Chutorian and Mohan Manubhai Trivedi. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):607–626, April 2009. 13, 89, 90, 93

[183] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010. 100

[184] N. Narayan, N. Sankaran, D. Arpit, K. Dantu, S. Setlur, and V. Govindaraju. Person re-identification for improved multi-person multi-camera tracking by continuous entity association. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 566–572, July 2017. 71

[185] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *Advances in Neural Information Processing Systems*, pages 2274–2284, 2017. 8

[186] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016. 8

[187] Anh Tuan Nghiem, Edouard Auvinet, and Jean Meunier. Head detection using kinect camera and its application to fall detection. In *Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on*, pages 164–169. IEEE, 2012. 12

[188] Aiden Nibali, Zhen He, Stuart Morgan, and Luke Prendergast. 3d human pose estimation with 2d marginal heatmaps. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019. 10

[189] Bruce Xiaohan Nie, Ping Wei, and Song-Chun Zhu. Monocular 3d human pose estimation by predicting depth on joints. In *International Conference on Computer Vision (ICCV)*, 2017. 10

[190] Jesus Nuevo, Luis M. Bergasa, and Pedro Jiménez. Rsmat: Robust simultaneous modeling and tracking. *Pattern Recognition Letters*, 31:2455–2463, December 2010. 102

[191] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *International Conference on 3D Vision (3DV)*, 2018. 10

[192] Ron MG op het Veld, RGJ Wijnhoven, Y Bondarev, et al. Detection and handling of occlusion in an object detection system. In *Video Surveillance and Transportation Imaging Applications 2015*, volume 9407, page 94070N. International Society for Optics and Photonics, 2015. 71

[193] Margarita Osadchy, Yann Le Cun, and Matthew L Miller. Synergistic face detection and pose estimation with energy-based models. *Journal of Machine Learning Research*, 8(May):1197–1215, 2007. 12

[194] Wanli Ouyang, Xingyu Zeng, and Xiaogang Wang. Partial occlusion handling in pedestrian detection with a deep model. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(11):2123–2137, 2016. 71

[195] Pashalis Padeleris, Xenophon Zabulis, and Antonis A Argyros. Head pose estimation on depth data based on particle swarm optimization. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, pages 42–49, 2012. 13, 97

[196] J. Pan and B. Hu. Robust occlusion handling in object tracking. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 71

[197] Georgios Th Papadopoulos, Apostolos Axenopoulos, and Petros Daras. Real-time skeleton-tracking-based human action recognition using kinect data. In *International Conference on Multimedia Modeling*, pages 473–483, 2014. 104

[198] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multiperson pose estimation in the wild. *arXiv preprint arXiv:1701.01779*, 8, 2017. 8, 9

[199] Chavdar Papazov, Tim K Marks, and Michael Jones. Real-time 3d head pose and facial landmark estimation from depth images using triangular surface patch features. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 4722–4730, 2015. 13, 97

[200] Eunbyung Park, Xufeng Han, Tamara L Berg, and Alexander C Berg. Combining multiple sources of knowledge in deep cnns for action recognition. In *Proc. IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8. IEEE, 2016. 101, 108

[201] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015. 121

[202] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016. 16, 98

[203] Abhishek Patil, Srikanth Malla, Haiming Gang, and Yi-Ting Chen. The h3d dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes. 2019. 6

[204] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7307–7316, 2018. 10

[205] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 10, 11, 36, 37, 39, 40

[206] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 10

[207] Tomas Pfister, James Charles, and Andrew Zisserman. Flowing convnets for human pose estimation in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1913–1921, 2015. 9, 39

[208] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4929–4937, 2016. 8, 9

[209] Alin-Ionut Popa, Mihai Zanfir, and Cristian Sminchisescu. Deep multitask architecture for integrated 2d and 3d human sensing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 47, 48

[210] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 15, 78, 79, 98, 99, 112

[211] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 7

[212] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 7, 43, 46, 49

[213] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016. 16

[214] Scott Reed, Aäron van den Oord, Nal Kalchbrenner, Victor Bapst, Matt Botvinick, and Nando de Freitas. Generating interpretable images with controllable structure. 2016. 16

[215] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. In *Advances in Neural Information Processing Systems*, pages 217–225, 2016. 16

[216] Ahmed Rekik, Achraf Ben-Hamadou, and Walid Mahdi. 3d face pose tracking using low quality depth cameras. In *VISAPP (2)*, pages 223–228, 2013. 11, 14, 97

[217] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 2015. 7, 8

[218] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014. 15

[219] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3d human pose estimation. *European Conference on Computer Vision (ECCV)*, 2018. 10

[220] Helge Rhodin, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua. Learning monocular 3d human pose estimation from multi-view images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 10

[221] Stephan R Richter, Zeeshan Hayder, and Vladlen Koltun. Playing for benchmarks. In *IEEE International Conference on Computer Vision*, 2017. 6, 7

[222] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European Conference on Computer Vision*, 2016. 7

[223] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*, 2016. 6

[224] Iasonas Kokkinos Riza Alp Guler, Natalia Neverova. Densepose: Dense human pose estimation in the wild. 2018. 71

[225] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 49

[226] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 10, 11, 46, 49

[227] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 351 of *LNCS*, pages 234–241. Springer, 2015. 78

[228] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 99, 106, 116

[229] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2016. 7

[230] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015. 5

[231] A. Sadeghian, A. Alahi, and S. Savarese. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 300–311, Oct 2017. 9, 33, 34

[232] Anwar Saeed and Ayoub Al-Hamadi. Boosted human head pose estimation using kinect camera. In *IEEE International Conference on Image Processing*, pages 1752–1756, 2015. 14, 97, 110

[233] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016. 82, 100

[234] Ricardo Sanchez-Matilla, Fabio Poiesi, and Andrea Cavallaro. Online multi-target tracking with strong and weak detections. In Gang Hua and Hervé Jégou, editors, *Computer Vision – ECCV 2016 Workshops*, pages 84–99, Cham, 2016. Springer International Publishing. 9, 33

[235] Benjamin Sapp, David Weiss, and Ben Taskar. Parsing human motion with stretchable models. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1281–1288. IEEE, 2011. 19

[236] István Sárándi, Timm Linder, Kai O Arras, and Bastian Leibe. Synthetic occlusion augmentation with volumetric heatmaps for the 2018 eccv posetrack challenge on 3d human pose estimation. *European Conference on Computer Vision (ECCV) - Workshops*, 2018. 10, 36

[237] Hamed Sarbolandi, Damien Lefloch, and Andreas Kolb. Kinect range sensing: Structured-light versus time-of-flight kinect. *Computer Vision and Image Understanding*, 139:1–20, 2015. 104

[238] Matthias Scholz, Martin Fraunholz, and Joachim Selbig. Nonlinear principal component analysis: neural network models and applications. In *Principal Manifolds for Data Visualization and Dimension Reduction*, 2008. 36

[239] Edgar Seemann, Kai Nickel, and Rainer Stiefelhagen. Head pose estimation using stereo vision for human-robot interaction. In *Proc. of Sixth International Conference on Face and Gesture Recognition*, pages 626–631. IEEE Computer Society, 2004. 11, 14

[240] Toby Sharp, Cem Keskin, Duncan Robertson, Jonathan Taylor, Jamie Shotton, David Kim, Christoph Rhemann, Ido Leichter, Alon Vinnikov, Yichen Wei, et al. Accurate, robust, and flexible real-time hand tracking. 2015. 7

[241] Lu Sheng, Jianfei Cai, Tat-Jen Cham, Vladimir Pavlovic, and King Ngi Ngan. A generative model for depth-based robust 3d facial pose tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4488–4497, 2017. 13, 97

[242] Jamie Shotton, Ross Girshick, Andrew Fitzgibbon, Toby Sharp, Mat Cook, Mark Finocchio, Richard Moore, Pushmeet Kohli, Antonio Criminisi, Alex Kipman, et al. Efficient human pose estimation from single depth images. 2012. 7

[243] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124, 2013. 94

[244] Tomas Simon, Hanbyul Joo, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. *CVPR*, 2017. 44

[245] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 21

[246] Monit Shah Singh, Vinaychandran Pondenkandath, Bo Zhou, Paul Lukowicz, and Marcus Liwicki. Transforming sensor data to the image domain for deep learning-an application to footstep detection. 111

[247] F. Solera, S. Calderara, and R. Cucchiara. Towards the evaluation of reproducible robustness in tracking-by-detection. In *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, Aug 2015. 18, 32

[248] Markus Storer, Martin Urschler, and Horst Bischof. 3d-mam: 3d morphable appearance model for efficient fine head pose estimation from still images. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 192–199. IEEE, 2009. 12

[249] Patrick Sudowe, Hannah Spitzer, and Bastian Leibe. Person attribute recognition with a jointly-trained holistic cnn model. In *ICCV Workshops*, pages 329–337. IEEE Computer Society, 2015. 15, 75

[250] Xiao Sun, Bin Xiao, Shuang Liang, and Yichen Wei. Integral human pose regression. *European Conference on Computer Vision (ECCV)*, 2018. 10, 36

[251] Yi Sun and Lijun Yin. Automatic pose estimation of 3d facial models. In *International Conference on Pattern Recognition*, pages 1–4, 2008. 12

[252] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 41, 44, 50

[253] Wei Tang, Pei Yu, and Ying Wu. Deeply learned compositional models for human pose estimation. In *European Conference on Computer Vision (ECCV)*, 2018. 36

[254] Geoffrey R Taylor, Andrew J Chosak, and Paul C Brewer. Ovvv: Using virtual worlds to design and evaluate surveillance systems. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2007. 7

[255] Bugra Tekin, Pablo Marquez Neila, Mathieu Salzmann, and Pascal Fua. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In *International Conference on Computer Vision (ICCV)*, 2017. 10

[256] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in neural information processing systems*, 2014. 39

[257] Róbert Torfason, Fabian Mentzer, Eiríkur Ágústsson, Michael Tschannen, Radu Timofte, and Luc Van Gool. Towards image understanding from deep compression without decoding. In *International Conference on Learning Representations*, 2018. 36

[258] Cuong Tran and Mohan Manubhai Trivedi. Vision for driver assistance: Looking at people in a vehicle. In *Visual Analysis of Humans*, pages 597–614. Springer, 2011. 90

[259] Mohan M Trivedi, Shinko Yuanhsien Cheng, Edwin MC Childers, and Stephen J Krotosky. Occupant posture analysis with stereo and thermal

infrared video: Algorithms and experimental evaluation. *IEEE Transactions on Vehicular Technology*, 53(6):1698–1712, 2004. 90

[260] Sergey Tulyakov, Radu-Laurentiu Vieriu, Stanislau Semeniuta, and Nicu Sebe. Robust real-time extreme head pose estimation. In *International Conference on Pattern Recognition*, pages 2263–2268, 2014. 11, 14

[261] Teodora Vatahska, Maren Bennewitz, and Sven Behnke. Feature-based head pose estimation from images. In *Proc. of 7th IEEE-RAS International Conference on Humanoid Robots*, pages 330–335, 2007. 12

[262] Paul Viola and Michael J Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004. 11

[263] Bingjie Wang, Wei Liang, Yucheng Wang, and Yan Liang. Head pose estimation with combined 2d sift and 3d hog features. In *Image and Graphics (ICIG), 2013 Seventh International Conference on*, pages 650–655. IEEE, 2013. 97

[264] Chaoyue Wang, Chang Xu, Chaohui Wang, and Dacheng Tao. Perceptual adversarial networks for image-to-image transformation. *IEEE Transactions on Image Processing*, 27(8):4066–4079, 2018. 16

[265] Jing Wang, Haibo He, and Danil V Prokhorov. A folded neural network autoencoder for dimensionality reduction. *Procedia Computer Science*, 2012. 36

[266] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2019. 7

[267] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018. 36

[268] Xiaolong Wang and Abhinav Gupta. Generative image modeling using style and structure adversarial networks. In *European Conference on Computer Vision*, pages 318–335. Springer, 2016. 16

[269] Xin Wang, Zhiqiang Hou, Wangsheng Yu, Lei Pu, Zefenfen Jin, and Xianxiang Qin. Robust occlusion-aware part-based visual tracking with object scale adaptation. *Pattern Recognition*, 81:456 – 470, 2018. 71

[270] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016. 8

[271] Jacob Whitehill and Javier R Movellan. A discriminative approach to frame-by-frame head pose tracking. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–7. IEEE, 2008. 12

[272] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649, 2017. 9, 33, 34

[273] Yu Xiang, Wongun Choi, Yuanqing Lin, and Silvio Savarese. Subcategory-aware convolutional neural networks for object proposals and detection. In *IEEE winter conference on applications of computer vision (WACV)*, 2017. 7

[274] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *European Conference on Computer Vision (ECCV)*, 2018. 36, 41, 50

[275] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, 2010. 5

[276] Xiang Xu and Ioannis A Kakadiaris. Joint head pose estimation and face alignment framework using global and local cnn features. In *Proc. 12th IEEE Conference on Automatic Face and Gesture Recognition, Washington, DC*, volume 2, 2017. 13

[277] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*, pages 776–791. Springer, 2016. 16

[278] Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, and Hao Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 3, 2017. 16

[279] Jiaolong Yang, Wei Liang, and Yunde Jia. Face pose estimation with combined 2d and 3d hog features. In *International Conference on Pattern Recognition*, pages 2492–2495, 2012. 14, 97

[280] Jimei Yang, Scott E Reed, Ming-Hsuan Yang, and Honglak Lee. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. In *Advances in Neural Information Processing Systems*, pages 1099–1107, 2015. 16

[281] Ruigang Yang and Zhengyou Zhang. Model-based head pose tracking with stereovision. In *Proc. of IEEE International Conference on Automatic Face and Gesture Recognition*, pages 255–260, 2002. 12

[282] S. Yang, P. Luo, C. C. Loy, and X. Tang. Faceness-net: Face detection through deep facial part responses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2017. 11

[283] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 10

[284] Hashim Yasin, Umar Iqbal, Bjorn Kruger, Andreas Weber, and Juergen Gall. A dual-source approach for 3d pose estimation from a single image. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 10

[285] Raymond A Yeh, Chen Chen, Teck Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5485–5493, 2017. 16

[286] Junho Yim, Heechul Jung, ByungIn Yoo, Changkyu Choi, Dusik Park, and Junmo Kim. Rotating your face using multi-task deep neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 676–684, 2015. 16

[287] Fengwei Yu, Wenbo Li, Quanquan Li, Yu Liu, Xiaohua Shi, and Junjie Yan. Poi: Multiple object tracking with high performance detection and appearance feature. In Gang Hua and Hervé Jégou, editors, *Computer Vision*

– *ECCV 2016 Workshops*, pages 36–42, Cham, 2016. Springer International Publishing. 9, 33, 34

[288] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 60

[289] Kevan Yuen, Sujitha Martin, and Mohan M Trivedi. On looking at faces in an automobile: Issues, algorithms and evaluation on naturalistic driving dataset. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 2777–2782. IEEE, 2016. 102

[290] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. 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)*, 2018. 10, 11, 47, 48

[291] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu. Deep network for the integrated 3d sensing of multiple people in natural images. In *Advances in Neural Information Processing Systems*, 2018. 11, 47, 48

[292] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012. 106

[293] Dong Zhang and Mubarak Shah. Human pose estimation in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2012–2020, 2015. 9

[294] Kaipeng Zhang, Lianzhi Tan, Zhifeng Li, and Yu Qiao. Gender and smile classification using deep convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–38, 2016. 14

[295] Ning Zhang, Manohar Paluri, Marc'Aurelio Ranzato, Trevor Darrell, and Lubomir Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1637–1644, 2014. 15

[296] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2248–2255, 2013. 18, 19

[297] Wuming Zhang, Zhixin Shu, Dimitris Samaras, and Liming Chen. Improving heterogeneous face recognition with conditional adversarial networks. *arXiv preprint arXiv:1709.02848*, 2017. 16

[298] Zichao Zhang, Henri Rebecq, Christian Forster, and Davide Scaramuzza. Benefit of large field-of-view cameras for visual odometry. 2016. 6

[299] Bo Zhao, Xiao Wu, Zhi-Qi Cheng, Hao Liu, Zequn Jie, and Jiashi Feng. Multi-view image generation from a single-view. *arXiv preprint arXiv:1704.04886*, 2017. 16

[300] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision*, 2015. 6

[301] Yang Zhong, Josephine Sullivan, and Haibo Li. Leveraging mid-level deep representations for predicting face attributes in the wild. In *IEEE International Conference on Image Processing*, pages 3239–3243, 2016. 14

[302] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 5

[303] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *International Conference on Computer Vision (ICCV)*, 2017. 10

[304] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv:1904.07850*, 2019. 7, 60

[305] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krahenbuhl. Bottom-up object detection by grouping extreme and center points. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 8

[306] Jianqing Zhu, Shengcai Liao, Zhen Lei, Dong Yi, and Stan Li. Pedestrian attribute classification in surveillance: Database and evaluation. In *IEEE International Conference on Computer Vision Workshops*, pages 331–338, 2013. 14

[307] Jianqing Zhu, Shengcai Liao, Dong Yi, Zhen Lei, and Stan Z Li. Multi-label cnn based pedestrian attribute learning for soft biometrics. In *International Conference on Biometrics*, pages 535–540, 2015. 15

[308] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face Alignment Across Large Poses: A 3D Solution. *ArXiv e-prints*, November 2015. 13