Matching and Recovering 3D People from Multiple Views

Alejandro Perez-Yus Antonio Agudo
Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Spain

Abstract

This paper introduces an approach to simultaneously match and recover 3D people from multiple calibrated cameras. To this end, we present an affinity measure between 2D detections across different views that enforces an uncertainty geometric consistency. This similarity is then exploited by a novel multi-view matching algorithm to cluster the detections, being robust against partial observations as well as bad detections and without assuming any prior about the number of people in the scene. After that, the multi-view correspondences are used in order to efficiently infer the 3D pose of each body by means of a 3D pictorial structure model in combination with physico-geometric constraints. Our algorithm is thoroughly evaluated on challenging scenarios where several human bodies are performing different activities which involve complex motions, producing large occlusions in some views and noisy observations. We outperform state-of-the-art results in terms of matching and 3D reconstruction.

1. Introduction

Human pose estimation is a quintessential computer vision problem which is experiencing a growing interest in fields such as sports, surveillance, activity recognition or motion capture. With the help of deep learning, astounding advances have been achieved in 2D [12, 23, 40]. Nevertheless, recovering poses in 3D is still an unsolved issue, especially in crowded scenes. In this work, we address the problem of 3D multi-body pose estimation from multiple calibrated views, which is a fairly common set-up in some of the aforementioned applications. While remarkable results have been achieved in multi-view reconstruction for a single body, the multi-body case represents a more challenging setting. In real applications, a set of people can move, deform, or even interact between them, producing complex motions which involve significant occlusions in some views. In addition, the number of people in the scene as well as their appearance are normally unknown.

A common way to tackle this problem is to split it in two stages. In the first one, a 2D pose detector is applied in every view to obtain body locations, which are then used in a posterior stage to infer the 3D pose. Nowadays there is a large collection of accurate methods that could be used for the first stage [16, 18]. However, to infer the 3D location of the body joints it is necessary to associate the detections across views as well as to the body they belong to. In this context, a typical approach to obtain correspondences is to use the epipolar constraint for each pair of views in its different variants. Unfortunately, this constraint could not be enough due to 2D noisy observations and artifacts such as occlusions. Moreover, trying to match each pair of images separately may produce inconsistency fails. Some works have tried to directly solve the association problem along with the 3D inference, by reasoning about all hypotheses in 3D that are geometrically consistent with 2D detections [6, 7, 21, 30]. These approaches are based on a 3D pictorial structure (3DPS) model, where the 3D body can be treated as an undirected graph that allows the inclusion of additional priors. Although these approaches can produce good results, they normally are very inefficient in terms of computational cost. More recently, other works have suggested to match the detected 2D poses among multiple views at the body level [20], and then inferring the 3D pose in a reduced state space, decreasing drastically the computational cost without sacrificing the accuracy.

We now move a step forward and tackle the problem of jointly matching and recovering 3D people from multiple views. Our approach exploits a set of 2D detections across views to group them in different 2D poses of the same body. To achieve that, we propose a robust multi-view matching algorithm that uses affinities at the body level based on an uncertainty geometric consistency, while it is robust to bad detections, noisy observations and occlusions. Then, the 3D pose of each body is efficiently inferred by using a 3DPS model with physico-geometric constraints. A pipeline of the whole method can be observed in Fig. 1. Furthermore, the accuracy of the matching and the 3D reconstruction we obtain improve those of state-of-the-art approaches.

2. Related Work

The problem of 3D human pose estimation from multiple views is a challenging task and covers many different areas

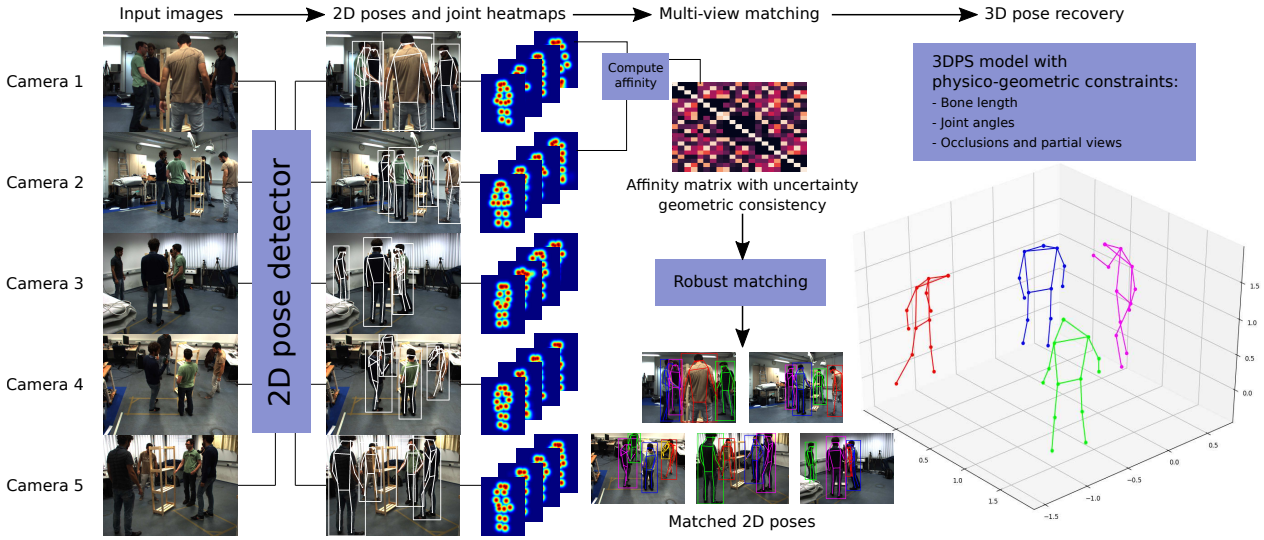

Figure 1. **Overview of our approach to infer 3D people from multiple views.** A set of calibrated RGB cameras from different point of views are observing an unknown number of people that can move, deform, and even interact between them producing large occlusions in some pictures. Our algorithm first obtains corrupted 2D pose detections in form of bounding boxes as well as the corresponding 2D joint positions and heatmaps. This information is used to compute an affinity matrix with uncertainty geometric consistency to later group bounding boxes by means of a robust multi-view matching algorithm. Once the bounding boxes of the same person in different images are grouped, *i.e.*, joint correspondences are solved, the physico-geometric 3D pose of every human person is recovered. Best viewed in color.

of knowledge in computer vision and learning. The starting point of any method usually is the estimation of 2D body poses for each separate view, which nowadays achieves incredible results [40, 50, 57]. For single-human pose detection, a basic approach is to predict the position of the body joints in the image, which usually comes in the form of a heatmap for each joint [15, 39, 52, 56]. In contrast, dealing with 2D pose estimation for multiple people requires more sophisticated methods, which can be divided in two general categories of approaches. On the one hand, there are top-down approaches [16, 23, 25, 36, 46, 53], which first detect the people in the scene and then apply a single-human pose estimator for each detection separately. Later, these solutions were extended to jointly incorporate the tracking and improve the results over time [19, 22, 54]. On the other hand, bottom-up approaches strive to extract all body joints in the image at once, as well as the associations between them to find human detections [12, 18, 26, 27, 34, 35, 45]. Other solutions like [32, 41] combined key-point detection with human segmentation. In general terms, bottom-up approaches are faster –even real-time [12]–, since they only need to process the image once, but they are normally less accurate than top-down approaches.

There are methods that go beyond 2D pose detection and infer the 3D pose from single images, either lifting the detected 2D poses into 3D [38, 44] or directly regressing 3D poses [42, 47, 49, 58]. Other approaches have directly inferred the multi-body 3D poses from a sequence of RGB images in the context of non-rigid structure from motion [1, 3]. However, the reconstruction accuracy of these approaches is not comparable with that based on multi-view. Facing the problem of pose recovery from multiple views allows to estimate accurate 3D poses with metric distances by combining deep-learning techniques with multiple view geometry. In [30] it was proposed a voxel grid discretization of the space and then uses the scores of the 2D part detector and reprojection error to obtain 3D joint positions, which then associates to different humans regarding the distance to the head point. Similarly, in [21] was projected the 2D score maps from the 2D pose detector to a shared 3D search space for clustering into different individuals. Other works exploited volumetric triangulation [29] to infer the 3D, or directly used the epipolar geometry [33].

In addition to multi-view geometry, most previous works are based on 3DPS [5, 6, 8, 11] in which nodes represent 3D locations of body joints and edges code pairwise relations between them. These works combined the confidence of the part detector together with some geometrical constraints. More recently, 3DPS-based models were used to train a network to infer the 3D pose from a single image [43]. These 3DPS approaches are often combined with matching strategies, where before inferring the 3D pose, a multi-view matching algorithm was performed to group the bodies across views, like [20], that uses geometry and appearance, or [13], that performs people matching with feet assignment. While these approaches are promising, the solution can still fail in difficult scenarios where the quality of the observations is not good. To improve upon that, there

are works that leverage temporal consistency [2, 14, 48] as an additional prior. Recent works are introducing neural networks to address the whole pipeline, such as [24, 51].

Our Contributions. We depart from previous work in that our solution simultaneously matches and recovers 3D people from multiple views. We can tackle scenarios with complex motions, where multiple people are performing different activities while producing body-location patterns with a high degree of overlapping. Our estimation is robust against bad detections and artifacts due to lack of visibility, self-occlusions, as well as noisy observations. To this end, we propose a novel affinity measure with uncertainty geometric consistency to match the 2D detections at the body level. After that, we group the 2D detections where every group contains the 2D poses of the same body in different views, thanks to our novel robust multi-view matching algorithm. Finally, we infer the 3D pose for every body separately by enforcing physico-geometric constraints.

3. Multi-view Matching

In this section, we propose our novel approach to match correspondences of human observations from multiple views. To this end, we first apply a 2D human pose detector. We have used the top-down method proposed by [16], even though our method is potentially applicable with any other. This detector provides a set of bounding boxes per image, and estimates the 2D position of each body joint with an associated heatmap. Ideally, every bounding box correspond to a human person, but unfortunately, the detection algorithm can still provide bad detections. Our goal is to match the detected bounding boxes belonging to the same body across views even when some detections are corrupt, being robust against self-occlusions or lack of visibility, and without assuming any information about the number of people (some consistent and inconsistent correspondences are shown in Fig. 2-right). To solve this problem, we first compute an affinity matrix between bounding boxes by enforcing an uncertainty geometric consistency, which is later exploited by an optimization algorithm to infer the correspondences between bodies. Thanks to this type of matching, we implicitly solve also the 2D body correspondences along the views.

3.1. Problem Statement

Let us assume the scene is observed by C RGB cameras where b_c bounding boxes are detected in the image $\mathcal{I}_c$ for the c-th camera. For every pair of views (c, d), we can define a $b_c \times b_d$ affinity matrix $\mathbf{A}_{cd}$, whose elements indicate the affinity scores and they should have higher entries for pairs of bounding boxes of the same object. Our problem consists in estimating correspondences between bounding boxes by means of a $b_c \times b_d$ binary partial permutation matrix $\mathbf{X}_{cd}$. To this end, the partial permutation matrix $\mathbf{X}_{cd}\,\forall\{c, d\}$ has to maximize the corresponding affinities $\mathbf{A}_{cd}\,\forall\{c, d\}$ while enforcing a cycle consistency along the C views. Additionally, this process must be robust against bad detections, since some of the b_c bounding boxes may not correspond to a person (an example with yellow box is displayed in Fig. 2-right).

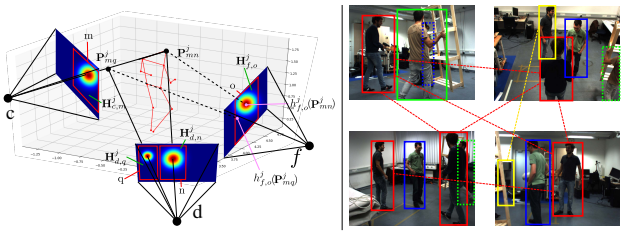

Figure 2. **Multi-view bounding-box correspondences**. **Left:** Affinity scores extraction from multiple views. To establish robust affinities between bounding boxes in views c and d, a novel view f can disambiguate the problem, voting for the bounding box n rather than the q as the best match for the bounding box m thanks to an uncertainty geometric consistency. **Right:** A real case with four views to be matched, and potential mistakes. Blue and green boxes show an example of consistent correspondences even for partial observations in some views (represented by dashed lines). Red boxes represent an example of inconsistent correspondences, since a bounding box is matched by two bounding boxes in one single image (see the case on the left/bottom part). Yellow boxes represent another example of inconsistent correspondences, since a bad detection is also matched. Best viewed in color.

3.2. Affinity Estimation

To compute affinities, let us consider a set of J body joint estimations for the b_c-th bounding box. From the human 2D pose detector, we can obtain for the j-th joint its 2D location as $\mathbf{p}^j_{c,b} = [x^j_{c,b}, y^j_{c,b}]^\top$. Similarly, we can denote as $\mathbf{H}^j_{c,b}$ the corresponding heatmap of $\mathbf{p}^j_{c,b}$, which consists of a matrix of the same size as the input image with the network prediction scores of the 2D joint position. Since the network is usually trained with 2D Gaussian profiles centered in the ground truth joint position (in our case, using the MSCOCO dataset [37]), the network predictions normally resemble a Gaussian, where higher entries imply closeness to the joint prediction and lower ones indicate farther distances. Next, we propose to exploit this observation to obtain the affinity scores between bounding boxes across views.

Affinity matrix from heatmaps. To compute the affinity matrix $\mathbf{A}_{cd}$ for views c and d, we use the multi-view projection matrices used to relate the 2D projection of a point with its corresponding 3D location, denoted $\mathbf{p}^j_{c,b} \equiv \pi_c(\mathbf{P}^j_{c,b})$, where π_c denotes an operator to perform projection in the c-th camera. Let m with $m \in \{1, \dots, b_c\}$ be a bounding box observed in the c-th camera, and n with $n \in \{1, \dots, b_d\}$ another bounding box in the d-th view. To obtain the affin-

ity score in the location $[m, n]$ within the matrix $\mathbf{A}_{cd}$, we proceed as follows. First, for every j-th point, we use the 2D locations $\mathbf{p}^j_{c,m}$ and $\mathbf{p}^j_{d,n}$ to virtually hallucinate a 3D point location that is projected back over the corresponding heatmaps. Once the point is projected, the scores $h^j_{c,m}$ and $h^j_{d,n}$ can be computed as:

$$h^j_{c,m} = \mathbf{H}^j_{c,m}(\pi_c(\mathcal{T}(\mathbf{p}^j_{c,m}, \mathbf{p}^j_{d,n}))), \quad (1)$$

$$h^j_{d,n} = \mathbf{H}^j_{d,n}(\pi_d(\mathcal{T}(\mathbf{p}^j_{c,m}, \mathbf{p}^j_{d,n}))), \quad (2)$$

where $\mathcal{T}(\cdot)$ denotes an operator to intersect rays. In practice, the virtual 3D point is minimizing the 3D distance to both rays. It is worth noting that if the re-projection lies out of the heatmap $\mathbf{H}^j_{c,m}$, for instance, the corresponding affinity score $h^j_{c,m}$ will be null. Every affinity score is normalized with the maximum value of the heatmap, obtaining $\hat{h}^j_{c,m}$ and $\hat{h}^j_{d,n}$, that finally are combined over the J points in the bounding box as:

$$\mathbf{A}_{cd}[m, n] = \frac{1}{J}\sum_{j=1}^{J} \frac{\hat{h}^j_{c,m} + \hat{h}^j_{d,n}}{2}, \quad (3)$$

where every entry is a value between 0 and 1 due to the process of normalization.

Enforcing uncertainty geometric consistency. Until now, we have taken only into account how the projection of a point from two views is observed in those views. However, that two bounding boxes from different views have a high similarity does not always mean that the match is correct (see an example in Fig. 2-left). When the system is composed of more than two views $C > 2$, we can extend our previous approach to handle more information, allowing us to filter out a lot of false positives. Particularly, we can apply Eq. (1) and check if the re-projection in a new heatmap is consistent, increasing or decreasing the corresponding affinity score accordingly.

To this end, we have to consider a new f-th view with $f \notin \{c, d\}$, for the o-th bounding box with $o = \{1, \ldots, b_f\}$, obtaining a score $h^j_{f,o}$ as:

$$h^j_{f,o} = \mathbf{H}^j_{f,o}(\pi_f(\mathcal{T}(\mathbf{p}^j_{c,m}, \mathbf{p}^j_{d,n})). \quad (4)$$

This process is repeated for every bounding box in the f-th view, selecting the highest score, and normalizing it. For simplicity, we assume the highest score is taken from the o-th bounding box. However, since not all points are observed by all cameras, the virtual 3D point could not be reprojected in the image and then this camera should not be taken into account. When the point is within the image, the f-th view is included to a group $\mathcal{V}$, which contains the set of cameras that can observe the point. Considering that, instead of taking a simple average as in Eq. (3), we now consider the set of cameras $\mathcal{V}$ as:

$$\mathbf{A}^j_{cd}[m, n] = \frac{1}{J}\sum_{j=1}^{J} \frac{\hat{h}^j_{c,m} + \hat{h}^j_{d,n} + \sum_{f\in\mathcal{V}} \hat{h}^j_{f,o}}{2 + V}, \quad (5)$$

where V denotes the number of cameras in the set $\mathcal{V}$. Thanks to this incorporation of priors, a potential matching of bounding boxes that produces a low score in other views it will see how its affinity is reduced. And backwards, potential matches with high scores in other views will maintain their high affinity.

3.3. Bounding-box Matching

Let $B = \sum_{c=1}^{C} b_c$ be the total number of detected bounding boxes in all views. The full correspondences can be coded by a $B \times B$ matrix $\mathbf{X}$ as:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_{11} & \mathbf{X}_{12} & \ldots & \mathbf{X}_{1C} \\ \mathbf{X}_{21} & \ddots & \ldots & \mathbf{X}_{2C} \\ \vdots & \vdots & \mathbf{X}_{cd} & \vdots \\ \mathbf{X}_{C1} & \ldots & \ldots & \mathbf{X}_{CC} \end{bmatrix}, \quad (6)$$

where every block $\mathbf{X}_{cd}$ codes the correspondences in the views $\{c, d\}$. To constrain self-matching, every $\mathbf{X}_{cc}$ block should be identity. As it was reported by [20], if the correspondences are cycle consistent, this matrix will be semidefinite $\mathbf{X} \succeq 0$ and low-rank, *i.e.*, the matrix can be factorized as $\mathbf{X} = \mathbf{Q}\mathbf{Q}^\top$ being $\mathbf{Q}$ a $B \times R$ matrix. For clean detections, the value $R \equiv N$ represents the number of bodies N in the scene. However, in real scenarios bad detections can appear becoming to be this value $R \equiv N + D$, where D denotes the total number of bad detections. In any case, neither N nor D are known in advance, but we can directly enforce the low-rank constraint by means of a nuclear norm as a convex relaxation [17]. We also impose sparsity in $\mathbf{X}$ since at most one value per row and column in $\mathbf{X}_{cd}$ is non-null, by minimizing the sum of values in $\mathbf{X}$. Finally, our estimation has to be robust against noisy affinity scores due to random corruptions. To this end, we directly model a residual noise by a $B \times B$ matrix $\mathbf{E}$, and applying a l_1-norm for estimation, and defining by $\mathbf{W}$ the affinity matrix with clean scores. Considering all the terms, our cost energy $\mathcal{A}(\mathbf{X}, \mathbf{W}, \mathbf{E})$ can be written as:

$$-\sum_{c=1}^{C}\sum_{d=1}^{C}\langle \mathbf{W}_{cd}, \mathbf{X}_{cd}\rangle + \beta\langle \mathbf{1}_B, \mathbf{X}\rangle + \gamma\|\mathbf{X}\|_* + \lambda\|\mathbf{E}\|_1, \quad (7)$$

$$\text{subject to} \quad \mathbf{A} = \mathbf{W} + \mathbf{E}$$

where $\langle\cdot,\cdot\rangle$ indicates an inner product, $\mathbf{1}_B$ is a $B \times B$ matrix of ones, and $\|\cdot\|_*$ and $\|\cdot\|_1$ denote the nuclear and l_1 norms, respectively. $\mathbf{A}$ is the concatenation of all $\mathbf{A}_{cd}$, that were computed in section 3.2. $\{\beta, \gamma, \lambda\}$ represents the set of

penalty weights. To make the previous optimization problem tractable, we incorporate some additional constraints as it is standard in matching algorithms [55, 59], such that:

$$\mathbf{X}_{cc} = \mathbf{I}_{b_c}\,, \tag{8}$$

$$\mathbf{X}_{cd} = \mathbf{X}_{dc}^{\top}\,, \tag{9}$$

$$\mathbf{X} \in \mathbb{R}^{[0,1]}\,, \tag{10}$$

$$\mathbf{0}_{b_c} \leq \mathbf{X}_{cd}\mathbf{1}_{b_d} \leq \mathbf{1}_{b_c}\,, \tag{11}$$

$$\mathbf{0}_{b_d} \leq \mathbf{X}_{cd}^{\top}\mathbf{1}_{b_c} \leq \mathbf{1}_{b_d}\,, \tag{12}$$

where $\mathbf{I}_{b_c}$ is a $b_c \times b_c$ identity matrix, and $\mathbf{0}_{b_c}/\mathbf{1}_{b_c}$ are a vector of b_c zeros/ones, respectively. Equation (8) imposes self-matching between bounding boxes in the same view, Eq. (9) enforces a symmetry in $\mathbf{X}$, Eq. (10) constrains the values to be real in [0,1], and Eqs. (11)-(12) enforce the doubly stochastic constraints. The set of matrices to satisfy the previous constraints will be denoted by $\mathcal{C}$.

The optimization problem in Eq. (7) in combination with the constraints in Eqs. (8)-(9)-(10)-(11)-(12) is convex, and it can be solved by various methods. For efficiently, in this paper we adopt the alternating direction method of multipliers [9], writing the equivalent problem as:

$$\underset{\mathbf{X},\mathbf{J},\mathbf{W},\mathbf{E}}{\arg\min}\ \langle \beta\mathbf{1}_B - \mathbf{W}, \mathbf{X}\rangle + \gamma\|\mathbf{J}\|_* + \lambda\|\mathbf{E}\|_1 \tag{13}$$

$$\begin{aligned} \text{subject to}\quad & \mathbf{A} = \mathbf{W} + \mathbf{E} \\ & \mathbf{X} = \mathbf{J} \\ & \mathbf{X} \in \mathcal{C} \end{aligned}$$

where we introduce the auxiliary variable $\mathbf{J}$. The equivalent augmented Lagrange function to be solved is:

$$\begin{aligned} &\underset{\mathbf{X}\in\mathcal{C},\mathbf{J},\mathbf{W},\mathbf{E}}{\arg\min}\ \langle \beta\mathbf{1}_B - \mathbf{W}, \mathbf{X}\rangle + \gamma\|\mathbf{J}\|_* + \lambda\|\mathbf{E}\|_1 \\ &+ \langle \mathbf{L}_1, \mathbf{A} - \mathbf{W} - \mathbf{E}\rangle + \langle \mathbf{L}_2, \mathbf{X} - \mathbf{J}\rangle \\ &+ \frac{\alpha}{2}\|\mathbf{A} - \mathbf{W} - \mathbf{E}\|_F^2 + \frac{\alpha}{2}\|\mathbf{X} - \mathbf{J}\|_F^2\,, \end{aligned} \tag{14}$$

where $\{\mathbf{L}_1, \mathbf{L}_2\} \in \mathbb{R}^{B\times B}$ are the Lagrange multipliers, $\|\cdot\|_F$ denotes the Frobenius norm, and $\alpha > 1$ is a penalty coefficient to improve convergence. Primal and dual variables are alternatively updated in closed form until convergence, while keeping fixed the rest of variables. The weight coefficients are determined empirically as $\beta = 0.4$, $\gamma = 1.2$, and $\lambda = 1.4$; but kept constant in all experiments we describe later. After optimizing, the estimated permutation matrix $\mathbf{X}$ is quantized by a threshold of 0.5.

4. 3D Pose Recovery

Once the 2D poses of the same person in different views $\mathcal{I}_c$ are known, we solve the 3D reconstruction problem. To this end, we adopt the widely used 3DPS model, since thanks to its versatility, we can easily incorporate additional priors that produce more accurate solutions.

In this context, the human body is considered as an undirected graphical model, where the graph nodes represent the body joints (elbow, knee, etc.) and the edges are the body parts connecting joints, *e.g.*, lower arm, upper leg, and so on. Our aim is to retrieve a 3D body configuration $\mathbf{Y} = [\mathbf{y}^{1\top}, \mathbf{y}^{2\top}, \ldots, \mathbf{y}^{J\top}]^{\top}$ with $\mathbf{y}^j = [x^j, y^j, z^j]^{\top}$, maximizing a posterior distribution of 3D poses $p(\mathbf{Y}|\mathcal{I})$ as:

$$\begin{aligned} &\underbrace{\prod_{c=1}^{C}\prod_{j=1}^{J} p(\mathcal{I}_c|\pi_c(\mathbf{y}_c^j))}_{\text{data term}} \cdot \underbrace{\prod_{j=1}^{J}\prod_{c=1}^{C_{\mathbf{Y}}} p(\mathbf{H}_c^j|\mathcal{H})}_{\text{heatmap prior}} \\ &\cdot \underbrace{\prod_{(i,j)\in\mathcal{B}} p(\mathbf{y}^i, \mathbf{y}^j)}_{\text{bone length prior}} \cdot \underbrace{\prod_{(i,j,k)\in\mathcal{L}} p(\mathbf{y}^i, \mathbf{y}^j, \mathbf{y}^k)}_{\text{joint angle prior}}, \end{aligned} \tag{15}$$

where $C_{\mathbf{Y}}$ denotes a subset of views used to generate $\mathbf{Y}$, $\mathcal{B}$ indicates the set of pairs the joints that form a bone, and $\mathcal{L}$ is the set of joints that form a limb. The data term codes the likelihood $p(\mathcal{I}_c|\pi_c(\mathbf{y}_c^j))$ and it is given by the 2D heatmap provided by [16]. This term characterizes the 2D spatial distribution of each joint in each image.

The first prior term $p(\mathbf{H}_c^j|\mathcal{H})$ is to code the quality of the observations via their heatmaps $\mathbf{H}_c^j$, which implicitly penalizes deviations with respect to a reference Gaussian heatmap used for training, and that it is denoted by $\mathcal{H}$. In real applications, only accurate observations produce a Gaussian-like heatmap, while unfocused predictions are usually a matter of occlusions or bad detections. To provide a probability, we take the reference heatmap $\mathcal{H}$, that is computed based on the parameters of the ground truth used for training the detection network [16], according to our current image resolution. The probability of this term is one for clean observations, and less than one for noisy detections, after comparing Gaussian distributions via their standard deviations.

The second prior term $p(\mathbf{y}^i, \mathbf{y}^j)$ is to code the spatial structural dependency between the adjacent joints $\mathbf{y}^i$ and $\mathbf{y}^j$, which implicitly constraints the bone length between them. This term can be modeled by a Gaussian distribution $p(d^{ij}; \mu^{ij}, \sigma^{ij})$, where d^{ij} is the Euclidean distance between the joints $\mathbf{y}^i$ and $\mathbf{y}^j$. μ^{ij} and σ^{ij} represent the mean and standard deviation, respectively; that are learned from the Human3.6M dataset [28]. The probability of this term is one for $d^{ij} \equiv \mu^{ij}$, and decreases towards zero as the values move away from the mean according to σ^{ij}.

The last prior term $p(\mathbf{y}^i, \mathbf{y}^j, \mathbf{y}^k)$ is to provide physical consistency in limbs g^{ijk} composed of the joints $\mathbf{y}^i$, $\mathbf{y}^j$, and $\mathbf{y}^k$, respectively. This term implicitly imposes constraints in the angles that the bones physically can have, defining the set of these configurations by $\mathcal{F}$. To enforce this prior,

we use the dictionary proposed by [4], where some joint angle limits are established. In this case, we penalize 3D human poses that are not physically possible by means of a Kronecker's delta generalization function as:

$$p(\mathbf{y}^i, \mathbf{y}^j, \mathbf{y}^k) = \begin{cases} 1 & \text{if } \; g^{ijk} \in \mathcal{F} \\ 0 & \text{if } \; g^{ijk} \notin \mathcal{F} \end{cases} \quad . \tag{16}$$

Inference. The problem of estimating the 3D pose by maximizing $p(\mathbf{Y}|\mathcal{I})$ in Eq. (15) can become very complex in terms of computational cost as the state space dimension increases [11]. To solve this, some works have simplified the process by setting the state space for each 3D joint to be the 3D proposals triangulated from all pairs of corresponding 2D joints [20], which are the only candidate points used to find the solution. However, this approach does not consider the accuracy gains that may come with having more than two views. Here, we additionally include a set of interpolated candidates from the initial proposals, obtained as weighted centroids –via the heatmap quality term– from different combinations of candidate points. As the number of views increases, in order to keep the computational cost low, the candidate points with small data term are discarded from the optimization at the beginning. Doing that, the complexity is drastically reduced while being able to achieve more accurate results at the same time.

5. Experimental Evaluation

We now provide our experimental evaluation, with quantitative results in terms of body matching and 3D pose estimation. To this end, we mainly use the datasets *Campus* and *Shelf* [8], but also show some qualitative and quantitative results from *KTH Football II* [31] and *CMU Panoptic* [30]. Additionally, we show quantitative comparisons with respect to competing techniques for all datasets. In the supplementary material, a video shows the full sequences.

5.1. Multi-Body Matching Evaluation

We first evaluate our approach in terms of multi-body 2D matching. Unfortunately, no ground truth is provided in the datasets we use for quantitative evaluation. To solve that, we run the 2D part detector [16] over the datasets *Campus* and *Shelf*, and then the bounding boxes from different views that correspond to the same person are manually annotated. It is worth noting that we match all human bounding boxes, including partial views of the people due to occlusions. Once the ground truth is available, we perform a precision/recall analysis to find true and false positives. As our algorithm to solve multi-body matching exploits an affinity matrix by means of an optimization algorithm, we also present an ablation study to show the effectiveness of every part. To make the analysis more complete, we include the partial affinities based on *Geometry*, appearance cues *(ReID)* and the combination of both, as proposed by [20], as well as their optimization algorithm.

		Campus			Shelf		
Affinity	Optimization	Precision	Recall	F1-score	Precision	Recall	F1-score
Geometry	[20]	95.72	88.03	90.49	94.61	92.56	93.40
ReID	[20]	98.86	90.90	94.10	90.67	74.34	80.79
Geom. + ReID [20]	[20]	99.31	93.27	95.71	97.19	87.62	91.82
Geometry	Ours	93.23	94.34	93.46	93.61	91.63	92.40
ReID	Ours	98.26	92.97	95.04	90.48	76.14	81.92
Geom. + ReID [20]	Ours	99.57	95.86	97.40	96.28	89.27	92.39
Ours	[20]	**99.94**	97.83	98.71	**99.71**	91.34	94.95
Ours	Ours	**99.94**	**98.56**	**99.13**	99.35	**94.22**	**96.53**

Table 1. **Quantitative evaluation on multi-body matching.** Multiple combinations are provided by considering the partial affinities geometry and re-identification (ReID), as well as their combination as it was done by [20]; and our proposal. Moreover, it is also included the optimization algorithms provided by [20] and ours. The table reports precision, recall and F1-score for the datasets *Campus* and *Shelf*. In all cases, accuracy is in [%].

Table 1 summarizes these results. As it can be seen, our combination provides the best solutions in both datasets showing the superiority of our formulation, producing almost perfect solutions in *Campus*. While an affinity based on *ReID* can produce better solutions than those based on *Geometry*, this type of cues can fail in scenarios where the subjects wear similar clothes or look alike, as it occurs in *Shelf*. Thanks to our novel affinity in combination with the robust optimization, we can produce more robust and stable solutions than state-of-the-art approaches, being less likely to propose an erroneous match (high precision), while being able to obtain most of the possible matches (high recall). Moreover, our approach is able to find matches of isolated body parts, as it occurs when a head or the limbs are occluded (see some examples in Fig. 3, matches highlighted by green arrows), where the competing approaches fail; and to avoid bad detections for matching, as it is displayed in Fig. 3 by red arrows. As expected, incorrect matches will produce worse 3D reconstructions (see 3D poses in Fig. 3-right obtained by [20]).

5.2. Multi-body 3D Pose Recovery

We now evaluate our multi-body 3D pose estimation. For quantitative evaluation, we provide several metrics depending on the experiment. The most used one is the Percentage of Correctly estimated Parts (PCP) [11], that reports the percentage of successful estimations of body parts, *i.e.*, when the mean distance of the part joints is less than the length of the bone multiplied by a threshold ϕ. Most of approaches consider $\phi = 0.5$, even though we will also include a more restrictive metric with $\phi = 0.2$, denoting them as PCP5 and PCP2, respectively. As many other works, we also report the Mean Per Joint Position Error (Euclidean distance between 3D joints and ground truth) measured in mm. The MPJPE is only measured when the pose is correctly estimated (*i.e.*, the MPJPE is less than 500mm). To evaluate the missing or correctly estimated joints, we pro-

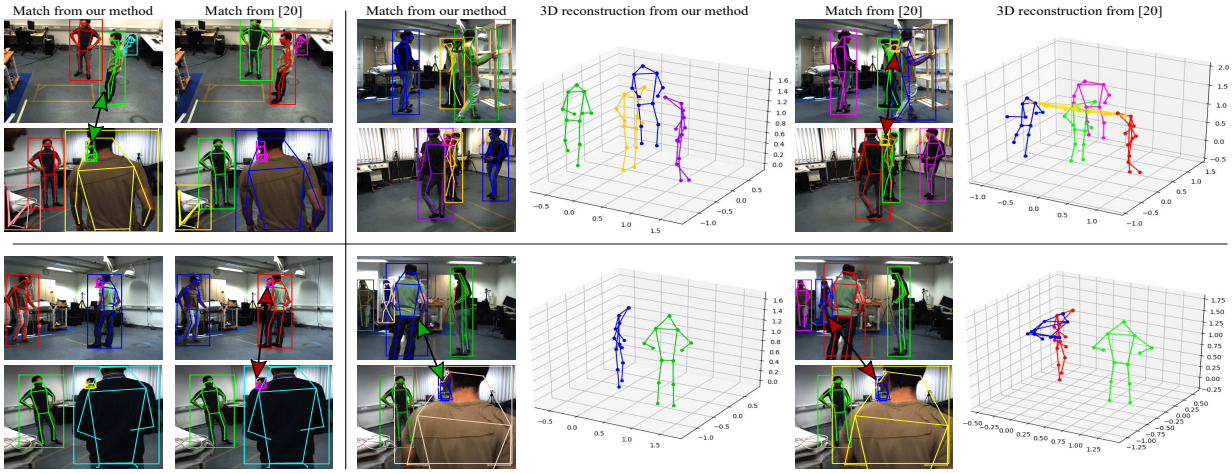

Figure 3. **Partial-view matchings. Left:** Multi-body matching in two instants (top and bottom, respectively) by using [20] and our algorithm. Each row corresponds to a different camera view in the same frame (for simplicity, we only represent two of five images per frame). Color bounding boxes and 2D poses show body associations across cameras. Our method can find correct matches even for strong occlusions (green double arrows), whereas avoid wrong matches obtained by competing approaches and highlighted by red double arrows. **Right:** Joint multi-body matching and 3D reconstruction in two novel instants by using [20] and our algorithm. Bad matchings can hallucinate incorrect poses. Best viewed in color.

	Campus			Shelf		
	PCP5	PCP2	MPJPE	PCP5	PCP2	MPJPE
Data term	95.88	56.64	79.5	96.40	64.82	58.1
+ bone length prior	95.94	57.48	80.5	96.23	64.53	58.5
+ heatmap prior	96.32	57.35	79.6	**96.63**	65.07	57.4
+ joint angle prior	95.80	57.59	80.6	96.39	64.86	58.1
+ all priors (Eq. (15))	96.19	58.25	79.4	96.41	64.70	58.0
+ all priors and interp.	**96.71**	**59.54**	**77.6**	96.53	**65.70**	**56.6**

Table 2. **Ablation study of our 3DPS-based algorithm.** The table shows the effect of every prior in Eq. (15) in combination with the data term. Last rows report the full model we use in Eq. (15), with and w/o the interpolation of new candidates in the inference. PCP5 and PCP2 are measured in percentage, and MPJPE in millimeters.

vide the 3D version of the Percentage of Correct Keypoints (PCK), which considers an estimated joint as correct when the distance w.r.t. the ground truth is within a certain threshold (we provide results with 50mm, 100mm and 150mm).

First, we evaluate the full energy in Eq. (15). To this end, we combine the data term with the priors one by one, to finally consider all, see Table 2. In *Campus*, a dataset with fewer views and farther distances, the bone length prior and joint angle priors are much more helpful to improve on accuracy due to enforcing physically-aware poses, as seen with the PCP2 metric. The heatmap prior improves on all PCP metrics in both datasets, showing the importance of removing inaccurate joints (*e.g.* due to occlusions, as it occurs in *Shelf*). The full energy generally improves on both datasets, particularly when the interpolation of candidates is also applied in the inference phase, which shows that simple triangulation does not provide the most accurate results.

In Table 3, we also compare our multi-body 3D reconstruction accuracy for previous datasets with respect to some baselines in literature. In this case, we provide the accuracy in terms of PCP5, as it was the only common metric in all those works. As it can be seen, our approach outperforms all methods in *Campus*, and achieves competitive solutions in *Shelf*. Some of the approaches here compared leverage temporal information, which we do not need in our method. Moreover, some other methods estimate the 3D pose using neural networks trained with the training frames from the target dataset, while our method is not personalized to any dataset in any way.

	Campus				Shelf			
	A1	A2	A3	Avg.	A1	A2	A3	Avg.
[6]	82.01	72.43	73.72	75.79	66.05	64.97	83.16	71.39
[5]	85.00	76.56	73.70	78.42	72.42	69.41	85.23	75.69
[7]	93.45	75.65	84.37	84.49	75.26	69.68	87.59	77.51
[21]	94.18	92.89	84.62	90.56	93.29	75.85	94.83	87.99
[10]	91.84	92.48	92.83	92.38	99.28	91.59	97.58	96.15
[20]	95.51	93.17	94.20	94.30	98.57	93.78	97.89	96.75
[48]	90.00	90.00	89.00	89.67	99.00	87.00	**98.00**	94.67
[48]$^+$	98.00	91.00	98.00	95.67	**99.80**	90.00	**98.00**	95.93
[14]$^+$	97.10	94.10	98.60	96.60	99.60	93.20	97.50	96.80
[51]*	97.60	93.80	**98.80**	96.70	99.30	94.10	97.60	97.00
[24]*	97.96	**94.81**	97.39	**96.71**	98.75	**96.22**	97.20	**97.39**
Ours	**98.37**	93.44	98.33	**96.71**	98.85	92.97	97.76	96.53

Table 3. **Quantitative evaluation and comparison on 3D pose estimation.** The table reports the 3D reconstruction accuracy in terms of PCP5 for actors {1,2,3} in datasets *Campus* and *Shelf*. The numbers are percentages. $^+$ includes temporal information, and * includes training from the target dataset.

Campus	PCP2 (1-3)	PCP2 (4)	MPJPE	PCK_{50}	PCK_{100}	PCK_{150}
[20]	56.95	-	79.5	25.99	72.25	91.79
Ours	**59.42**	-	**77.6**	**26.82**	**75.55**	**94.45**
Shelf	PCP2 (1-3)	PCP2 (4)	MPJPE	PCK_{50}	PCK_{100}	PCK_{150}
[20]	64.40	24.24	58.0	50.24	89.31	97.38
Ours	**65.70**	**35.15**	**57.0**	**51.27**	**90.48**	**97.80**

Table 4. **Quantitative evaluation on accurate metrics.** 3D reconstruction error in terms of an MPJPE, as well as the accuracy by using PCP2 (actors 1-3 and 4, respectively) and PCK (with 50, 100 and 150mm threshold) in datasets *Campus* and *Shelf*, compared to [20]. In millimeters and percentages, respectively.

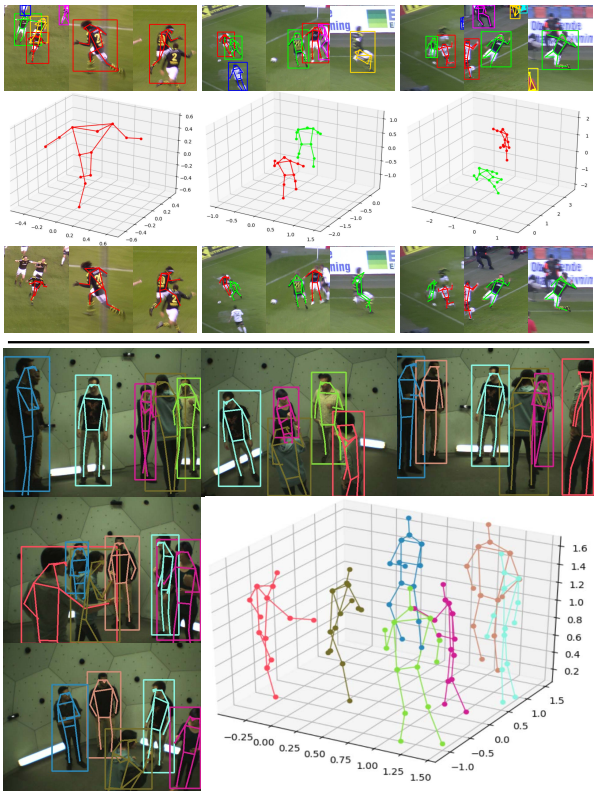

Figure 4. **Joint matching and 3D pose estimation in KTH Football II and CMU Panoptic datasets**. **Top:** Successful results on the *KTH Football II* dataset [31], where our method faces odd postures, motion blur, occlusions and high appearance similarity between players. For each frame, three views are shown from left to right, with the matched 2D detections on top and the 3D reprojection below. 3D reconstruction is shown in the middle. **Bottom:** One frame on the *CMU Panoptic* dataset [30], including matchings and 3D reconstruction from five cameras (up to seven people).

Additionally, we use the source code provided by [20] to report a full analysis, including the remaining metrics defined previously. These results are summarized in Table 4. As this analysis is performed from scratch, we also include the results for Actor 4 from *Shelf*, often omitted in the literature since it appears mostly occluded and it is not easy to recover its pose. We can observe our method clearly outperforms [20] in all metrics, meaning that it is much more accurate since estimated joints are closer to the ground truth, especially in highly occluded instances (*e.g.* Actor 4).

We now show some experiments in *KTH Football II* [31] and *CMU Panoptic* [30] datasets. First, soccer players move quickly and perform drastic motions at a very far distance, representing a challenging scenario for the matching algorithm due to motion blur. In the second one, the scenario is theoretically controlled but a large number of people appear, producing strong occlusions in many views, as it is shown in the images. As it can be seen in Fig. 4, our algorithm produces robust and physically-aware joint 2D/3D solutions even for strong occlusions and noisy observations.

Method	UA	LA	UL	LL	Average
[11]*	60.0	35.0	100.0	90.0	71.3
[31]	89.0	68.0	100.0	99.0	89.0
[6]	68.0	56.0	78.0	70.0	68.0
[7]	98.0	72.0	99.0	92.0	90.3
[21]	97.5	94.9	100.0	99.0	97.8
[48]	99.0	99.0	98.0	93.0	97.3
[48]$^{+}$	100.0	100.0	99.0	99.7	99.7
[43]	100.0	100.0	100.0	100.0	100.0
Ours	100.0	98.1	99.1	98.4	98.9
Ours (PCP2)	79.0	67.5	89.1	92.8	82.1

Table 5. **Quantitative evaluation on KTH Football II dataset** (Sequence 1, Player 2) with three cameras. Following previous works, we report the accuracy by PCP5 –excepting the last row– in percentage for some limbs: UA = upper arms, LA = lower arms, UL = upper legs, and LL = lower legs. * only uses several frames to compute PCP5. + includes temporal smoothing.

Quantitatively, we consider the sequence #1 of player #2 on the KTH Football II dataset [31] to compare to other methods. Our results are reported in Table 5. It is worth noting that our solution outperforms all comparable methods, and it is very close to that reported by [48], that used tracking of human poses and temporal smoothness priors, an extra information that is not required by our approach. We observe that [43] reported perfect results here, but it is important to point out that said method only works for one single human. It is, therefore, incapable of handling multiple humans as our method does, and thus it solves a different and more constrained problem needing strong assumptions. In contrast, our method is able to generalize to much more complex situations. In the supplementary material we show some additional quantitative results from CMU Panoptic.

6. Conclusion

In this paper we have presented a novel solution to jointly match and recover 3D people from multiple views. Our approach can efficiently handle noisy observations as well as cope with large occlusions, and without assuming any information about the number of people in the scene. For this purpose, we have proposed a strategy to define similarities between 2D detections by enforcing an uncertainty geometric consistency. This measure is then exploited by a robust multi-view matching algorithm that groups the detections in terms of body similarity. Once the correspondences are known, we apply a 3DPS-based algorithm to infer the 3D poses, enforcing physico-geometric constraints. We have thoroughly evaluated the approach on challenging scenarios involving interacting people performing complex motions. In the future we aim at extending our research to perform recognition of human activities.

Acknowledgments: This work has been partially supported by the Spanish Ministry of Science and Innovation under project MoHuCo PID2020-120049RB, and by the ERA-Net Chistera project IPALM PCI2019-103386.

References

[1] A. Agudo and F. Moreno-Noguer. DUST: Dual union of spatio-temporal subspaces for monocular multiple object 3D reconstruction. In *CVPR*, 2017.

[2] A. Agudo and F. Moreno-Noguer. Deformable motion 3D reconstruction by union of regularized subspaces. In *ICIP*, 2018.

[3] A. Agudo and F. Moreno-Noguer. Robust spatio-temporal clustering and reconstruction of multiple deformable bodies. *TPAMI*, 41(4):971–984, 2019.

[4] I. Akhter and M. J. Black. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *CVPR*, 2015.

[5] S. Amin, M. Andriluka, M. Rohrbach, and B. Schiele. Multi-view pictorial structures for 3D human pose estimation. In *BMVC*, 2013.

[6] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3D pictorial structures for multiple human pose estimation. In *CVPR*, 2014.

[7] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3D pictorial structures revisited: Multiple human pose estimation. *TPAMI*, 38(10):1929–1942, 2015.

[8] V. Belagiannis, X. Wang, B. Schiele, P. Fua, S. Ilic, and N. Navab. Multiple human pose estimation with temporally consistent 3D pictorial structures. In *ECCV*, 2014.

[9] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *FTML*, 3(1):1–122, 2011.

[10] L. Bridgeman, M. Volino, J-Y. Guillemaut, and A. Hilton. Multi-person 3D pose estimation and tracking in sports. In *CVPR*, 2019.

[11] M. Burenius, J. Sullivan, and S. Carlsson. 3D pictorial structures for multiple view articulated pose estimation. In *CVPR*, 2013.

[12] Z. Cao, T. Simon, S-E. Wei, and Y. Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *CVPR*, 2017.

[13] He Chen, Pengfei Guo, Pengfei Li, Gim Hee Lee, and Gregory Chirikjian. Multi-person 3d pose estimation in crowded scenes based on multi-view geometry. In *ECCV*, 2020.

[14] Long Chen, Haizhou Ai, Rui Chen, Zijie Zhuang, and Shuang Liu. Cross-view tracking for multi-human 3d pose estimation at over 100 fps. In *CVPR*, 2020.

[15] Y. Chen, C. Shen, X-S. Wei, L. Liu, and J. Yang. Adversarial posenet: A structure-aware convolutional network for human pose estimation. In *ICCV*, 2017.

[16] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun. Cascaded pyramid network for multi-person pose estimation. In *CVPR*, 2018.

[17] Y. Chen, H. Xu, C. Caramanis, and S. Sanghavi. Robust matrix completion with corrupted columns. In *ICML*, 2011.

[18] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *CVPR*, 2020.

[19] A. Doering, U. Iqbal, and J. Gall. Jointflow: Temporal flow fields for multi person pose tracking. In *BMVC*, 2018.

[20] J. Dong, W. Jiang, Q. Huang, H. Bao, and X. Zhou. Fast and robust multi-person 3D pose estimation from multiple views. In *CVPR*, 2019.

[21] S. Ershadi-Nasab, E. Noury, S. Kasaei, and E. Sanaei. Multiple human 3D pose estimation from multiview images. *Multimedia Tools and Applications*, 77(12):15573–15601, 2018.

[22] R. Girdhar, G. Gkioxari, L. Torresani, M. Paluri, and D. Tran. Detect-and-track: Efficient pose estimation in videos. In *CVPR*, 2018.

[23] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *ICCV*, 2017.

[24] Congzhentao Huang, Shuai Jiang, Yang Li, Ziyue Zhang, Jason Traish, Chen Deng, Sam Ferguson, and Richard Yi Da Xu. End-to-end dynamic matching network for multi-view multi-person 3d pose estimation. In *ECCV*, 2020.

[25] S. Huang, M. Gong, and D. Tao. A coarse-fine network for keypoint localization. In *ICCV*, 2017.

[26] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres, and B. Schiele. Arttrack: Articulated multi-person tracking in the wild. In *CVPR*, 2017.

[27] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV*, 2016.

[28] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *TPAMI*, 36(7):1325–1339, 2014.

[29] K. Iskakov, E. Burkov, V. Lempitsky, and Y. Malkov. Learnable triangulation of human pose. In *ICCV*, 2019.

[30] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. Godisart, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *TPAMI*, 41(1):190–204, 2017.

[31] V. Kazemi, M. Burenius, H. Azizpour, and J. Sullivan. Multi-view body part recognition with random forests. In *BMVC*, 2013.

[32] M. Kocabas, S. Karagoz, and E. Akbas. Multiposenet: Fast multi-person pose estimation using pose residual network. In *ECCV*, 2018.

[33] M. Kocabas, S. Karagoz, and E. Akbas. Self-supervised learning of 3D human pose using multi-view geometry. In *CVPR*, 2019.

[34] S. Kreiss, L. Bertoni, and A. Alahi. Pifpaf: Composite fields for human pose estimation. In *CVPR*, 2019.

[35] J. Li, W. Su, and Z. Wang. Simple pose: Rethinking and improving a bottom-up approach for multi-person pose estimation. In *AAAI*, 2020.

[36] J. Li, C. Wang, H. Zhu, Y. Mao, H-S. Fang, and C. Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *CVPR*, 2019.

[37] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[38] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3D human pose estimation. In *ICCV*, 2017.

[39] G. Moon, J. Y. Chang, and K. M. Lee. Posefix: Model-agnostic general human pose refinement network. In *CVPR*, 2019.

[40] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016.

[41] G. Papandreou, T. Zhu, L-C. Chen, S. Gidaris, J. Tompson, and K. Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *ECCV*, 2018.

[42] G. Pavlakos, X. Zhou, and K. Daniilidis. Ordinal depth supervision for 3D human pose estimation. In *CVPR*, 2018.

[43] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Harvesting multiple views for marker-less 3D human pose annotations. In *CVPR*, 2017.

[44] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*, 2019.

[45] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*, 2016.

[46] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.

[47] X. Sun, J. Shang, S. Liang, and Y. Wei. Compositional human pose regression. In *ICCV*, 2017.

[48] J. Tanke and J. Gall. Iterative greedy matching for 3D human pose tracking from multiple views. In *GCPR*, 2019.

[49] D. Tome, C. Russell, and L. Agapito. Lifting from the deep: Convolutional 3D pose estimation from a single image. In *CVPR*, 2017.

[50] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 2014.

[51] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In *ECCV*, 2020.

[52] S-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016.

[53] B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018.

[54] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu. Pose flow: Efficient online pose tracking. In *BMVC*, 2018.

[55] J. Yan, Y. Li, W. Liu, H. Zha, X. Yang, and S. M. Chu. Graduated consistency-regularized optimization for multi-graph matching. In *ECCV*, 2014.

[56] F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu. Distribution-aware coordinate representation for human pose estimation. In *CVPR*, 2020.

[57] F. Zhang, X. Zhu, and M. Ye. Fast human pose estimation. In *CVPR*, 2019.

[58] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei. Towards 3D human pose estimation in the wild: a weakly-supervised approach. In *ICCV*, 2017.

[59] X. Zhou, M. Zhu, and K. Daniilidis. Multi-image matching via fast alternating minimization. In *ICCV*, 2015.