DOPE: Distillation Of Part Experts for whole-body 3D pose estimation in the wild

Philippe Weinzaepfel Romain Brégier Hadrien Combaluzier
Vincent Leroy Grégory Rogez

NAVER LABS Europe

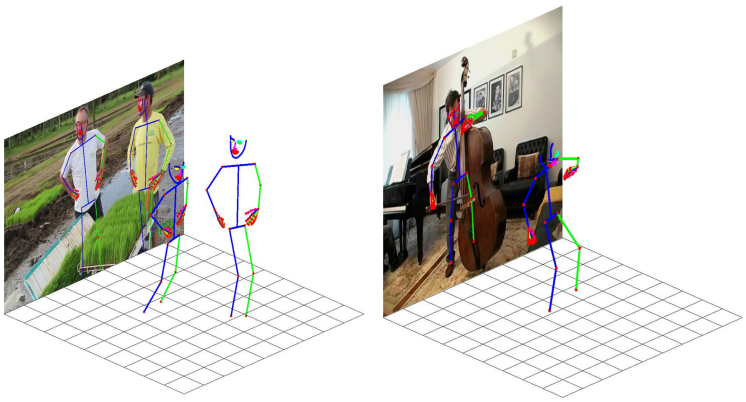


Fig. 1: Results of our DOPE approach for 2D-3D whole-body pose estimation.

Abstract. We introduce DOPE, the first method to detect and estimate whole-body 3D human poses, including bodies, hands and faces, in the wild. Achieving this level of details is key for a number of applications that require understanding the interactions of the people with each other or with the environment. The main challenge is the lack of in-the-wild data with labeled whole-body 3D poses. In previous work, training data has been annotated or generated for simpler tasks focusing on bodies, hands or faces separately. In this work, we propose to take advantage of these datasets to train independent experts for each part, namely a body, a hand and a face expert, and distill their knowledge into a single deep network designed for whole-body 2D-3D pose detection. In practice, given a training image with partial or no annotation, each part expert detects its subset of keypoints in 2D and 3D and the resulting estimations are combined to obtain whole-body pseudo ground-truth poses. A distillation loss encourages the whole-body predictions to mimic the experts' outputs. Our results show that this approach significantly outperforms the same whole-body model trained without distillation while staying close to the performance of the experts. Importantly, DOPE is computationally less demanding than the ensemble of experts and can achieve real-time performance. Test code and models are available at `https://europe.naverlabs.com/research/computer-vision/dope`.

Keywords: Human pose estimation, human pose detection, 3D pose estimation, 2D pose estimation, body pose estimation, hand pose estimation, face landmarks estimation

1 Introduction

Understanding humans in real-world images and videos has numerous potential applications ranging from avatar animation for augmented and virtual reality [15, 46] to robotics [21, 23]. To fully analyze the interactions of people with each other or with the environment, and to recognize their emotions or activities, a detailed pose of the whole human body would be beneficial. This includes 3D body keypoints, *i.e.*, torsos, arms and legs, that give information on the global posture of the persons, but also detailed information about hands and faces to fully capture their expressiveness. The task of whole-body 3D human pose estimation has been mainly addressed part by part as indicated by the large literature on estimating 3D body pose [3, 24, 38, 46, 48, 55], 3D hand pose [10, 49, 67, 73] or 3D face landmarks and shape [9, 57] in the wild. These methods now reach outstanding performances on their specific tasks, and combining them in an efficient way is an open problem.

More recently, a few approaches have been introduced that capture body, hands and face pose jointly. Hidalgo *et al.* [29] extend OpenPose [12] to predict 2D whole-body poses in natural images. To train their multi-task learning approach, they partly rely on datasets for which adding 2D pose annotations is possible, *e.g.*, adding 2D hand pose annotations [58] to the MPII body pose dataset [1]. Such annotation scheme is not possible when dealing with 3D poses. Importantly, they observe global failure cases when a significant part of the target person is occluded or outside of the image boundaries. Some other works have leveraged expressive parametric human models composed of body, hand and face components stitched together such as Adam [37, 64] or SMPL-X [51]. These optimization-based approaches remain sensitive to initialization and are usually slow to converge. Their performance highly depends on the intermediate estimation of the 3D orientations of body parts or 2D keypoint locations, and is therefore limited in cases of occlusions or truncations at the image boundary compared to more direct approaches.

In this paper, we propose the first learning-based method that, given an image, detects the people present in the scene and directly predicts the 2D and 3D poses of their bodies, hands and faces, see examples in Figure 1. Inspired by LCR-Net++ [55], a Faster R-CNN like architecture [53] tailored for in-the-wild 2D-3D body pose estimation, we design a classification-regression network where the object categories to detect are body, hand and face pose classes. In a second step, a class-specific regression is applied to refine body, hand and face pose estimates by deforming the average pose of each class both in 2D and 3D.

There exists no in-the-wild dataset to directly train our network, *i.e.*, images with 3D pose annotations for body, hand and face poses. Such data could only be obtained in specific controlled environments, *e.g.* in motion capture rooms or through computer-generation, which would not suit our purpose of whole-body pose estimation in unconstrained scenarios. However, multiple in-the-wild datasets are available for each independent task, *i.e.*, for 3D body pose estimation [38, 46], 3D hand pose estimation [25, 70, 73] or 3D facial landmark estimation [9, 18]. Task-specific methods trained on these datasets perform well in

Fig. 2: Overview of our DOPE training scheme. Each training image is processed by the part experts to detect their specific parts and estimate their 2D and 3D poses. The resulting detections are combined to obtain the whole-body poses used as ground-truth for this image when training our network. We show only 2D poses for the sake of clarity but we also distill the 3D poses.

practice but our experiments show that training our single model for whole-body 3D pose estimation on the union of these datasets leads to poor performances. Each dataset being annotated with partial pose information (*i.e.*, its specific part), unannotated parts are mistakenly considered as negatives by our detection framework, burdening the performance of the network.

To handle this problem, we propose to train independent experts for each part, namely body, hand and face experts, and distill their knowledge to our whole-body pose detection network designed to perform the three tasks jointly. In practice, given a training image with partial or no annotation, each part expert detects and estimates its subset of keypoints, in 2D and 3D, and the resulting estimations are combined to obtain whole-body pseudo ground-truth poses for the whole-body network. Figure 2 illustrates this training scheme. A distillation loss is applied on the network's output to keep it close to the experts' predictions. We name our method **DOPE** for **D**istillation **O**f **P**art **E**xperts. Our unified DOPE model performs on par with the part experts when evaluating each of the three tasks on dedicated datasets, while being computationally less demanding than the ensemble of experts and achieving real-time performances. In summary, we propose (a) a new architecture that can detect and estimate the whole-body 2D-3D pose of multiple people in the wild in real-time and (b) a novel and effective training scheme based on distillation that leverages previous data collection efforts for the individual subtasks.

This paper is organized as follows. After reviewing the related work (Section 2), we present our DOPE method in Section 3. Finally, experimental results for body, hand and face pose estimation are reported in Section 4.

2 Related work

The problem of 3D human whole-body pose estimation has been mainly tackled by breaking the body into parts and focusing on the pose inference of these parts separately. In the following, we briefly review the state of the art for each of these subtasks, before summarizing the few approaches that predict the 3D pose of the entire body, and finally discussing existing distillation methods.

3D body pose estimation. Two basic categories of work can be found in the recent literature: (a) approaches that directly estimate the 3D body keypoints from an input image [46, 48, 52, 54, 55] and (b) methods that leverage 2D human pose estimation [6,13,38,45]. The latter ones rely on a previous localization of the body keypoints in the image, through an off-the-shelf 2D pose detector [12,19,27], and lift them to 3D space [13, 45] or, as in [6, 38], use them to initialize the optimization procedure of a parametric model of the human body such as SMPL [43]. For our body expert, we employ LCR-Net++ [55] that jointly estimates 2D and 3D body poses from the image and has demonstrated robustness to challenging in-the-wild scenarios, *i.e.*, showing multiple interacting persons, with cluttered backgrounds and/or captured under severe occlusions and truncations.

3D hand pose estimation. 3D hand pose estimation from depth data has been studied for many years and state-of-the-art results on this task are now impressive as shown in a recent survey [60]. RGB-based 3D hand pose estimation is more challenging, and has gained interest in recent years. Regression-based techniques try to directly predict 3D location of hand keypoints [66] or even the vertices of a mesh [22] from an input image. Some methods incorporate priors by regressing parameters of a deformable hand model such as MANO [8, 26, 56], and many techniques leverage intermediate representations such as 2D keypoints heatmaps to perform 3D predictions [10, 49, 71, 73]. However, pose estimation is often performed on an image cropped around a single hand, and hand detection is performed independently. For our hand expert, we therefore use the detector of [55] (adapted to hands) that recently achieved outstanding performances in RGB-based 3D hand pose estimation under hand-object interaction [2].

3D face pose estimation. As with hands, the recovery of the pose of a face is typically performed from an image crop, by detecting particular 2D facial landmarks [63]. To better perceive the 3D pose and shape of a face, some works propose to fit a 3D Morphable Model [5, 7, 72] or to regress dense 3D face representations [17, 20, 34]. In this work, we also adopt [55] as face expert, resulting in an hybrid model-free approach that regresses 3D facial landmarks independently from their visibility, as in the approach introduced for the Menpo 3D benchmark [18].

3D Whole-body pose estimation. The few existing methods [51, 64] that predict the 3D pose of the whole-body all rely on parametric models of the human body, namely Adam [37] and SMPL-X [51]. These models are obtained by combining body, hand and face parametric models. Adam stitches together three different models: a simpler version of SMPL for the body, an artist-created rig for the hands, and the FaceWarehouse model [11] for the face. In the case of SMPL-X, the SMPL body model is augmented with the FLAME head model [40]

and MANO [56]. A more realistic model is obtained in the case of SMPL-X by learning the shape and pose-dependent blend shapes. Both methods are based on an optimization scheme guided by 2D joint locations or 3D part orientations. Monocular Total Capture [64] remains limited to a single person while for SMPL-X [51], the optimization strategy is applied independently on each person detected by OpenPose [12]. Optimizing over the parameters of such models can be time-consuming and the performance often depends on a correct initialization. Our approach is the first one that predicts whole-body 3D pose without relying on the optimization of a parametric model and can make real-time predictions of multiple 3D whole-body poses in real-world scenes. In addition, our DOPE training scheme can leverage datasets that do not contain ground-truth for all the parts at once.

Distillation. Our learning procedure is based on the concept of distillation which was proposed in the context of efficient neural network computation by using class probabilities of a higher-capacity model as soft targets of a smaller and faster model [30]. Distillation has been successfully employed for several problems in computer vision such as object detection [14], video classification [4], action recognition [16], multi-task learning [42] or lifelong learning [32]. In addition to training a compact model [4, 14], several works [16, 31] have shown that distillation can be combined with privileged information [61], also called generalized distillation [44] in order to train a network while leveraging extra modalities available for training, *e.g.* training on RGB and depth data while only RGB is available at test time. In this paper, we propose to use distillation in order to transfer the knowledge of several body-part experts into a unified network that outputs a more complete representation of the whole human body.

3 DOPE for 2D-3D whole-body pose estimation

After introducing our architecture for multi-person whole-body pose estimation (Section 3.1), we detail our training procedure based on distillation (Section 3.2).

3.1 Whole-body pose architecture

We propose a method that, given an image, detects the people present in the scene and directly predicts the 2D and 3D poses of their bodies, hands and faces. Our network architecture takes inspiration from [55], which extends a Faster R-CNN like architecture [53] to the problem of 2D-3D body pose estimation and has shown to be robust in the wild. We thus design a Localization-Classification-Regression network where the objects to be detected are bodies, hands and faces with respectively J_B, J_H and J_F keypoints to be estimated in 2D and 3D. Figure 3 shows an overview of this architecture.

Localization. Given an input image, convolutional features (ResNet50 [28] up to block3 in practice) are computed and fed into a Region Proposal Network (RPN) [53] to produce a list of candidate boxes containing potential body, hand or face instances. Although they might belong to the same person, we specifically

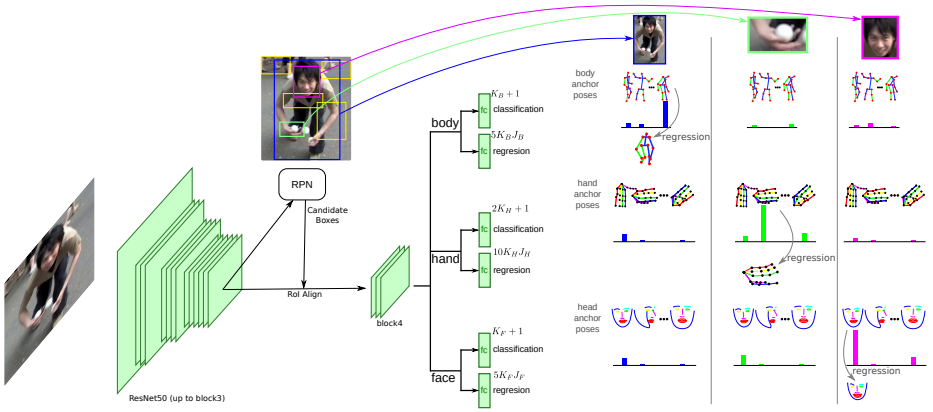


Fig. 3: Overview of our whole-body pose estimation architecture. Given an input image, convolutional features are computed and fed into a Region Proposal Network (RPN) to produce a list of candidate boxes. For each box, after RoI-Align and a few additional layers, 6 final outputs are computed (2 for each part). The first one returns a classification score for each anchor-pose corresponding to this part (including a background class not represented for clarity) while the second one returns refined 2D-3D pose estimates obtained through class-specific regression from the fitted anchor pose.

treat the parts as separate objects to be robust to cases where only a face, a hand or a body is visible in the image. Our network can also output whole-body poses of multiple people at once, when their different parts are visible. The candidate boxes generated by the RPN are used to pool convolutional features using RoI Align, and after a few additional layers (block4 from ResNet50 in practice), they are fed to the classification and regression branches, 6 in total: one classification and one regression branch per part.

Classification. Classification is performed for the three sub-tasks: body, hand and face classification. As in [55], pose classes are defined by clustering the 3D pose space. This clustering is applied independently in the 3 pose spaces, corresponding to the 3 parts, obtaining respectively a set of K_B, K_H and K_F classes for bodies, hands and faces. Note that to handle left and right hands with the same detector, we actually consider $2 \times K_H$ hand classes, K_H for each side. For each classification branch, we also consider an additional background class to use the classifier as a detector. Therefore, each candidate box is classified into $K_B + 1$ labels for body classes, $2K_H + 1$ for hands and $K_F + 1$ for faces.

Regression. In a third step, a class-specific regression is applied to estimate body, hand and face poses in 2D and 3D. First, for each class of each part, we define offline the 'anchor-poses', computed as the average 2D and 3D poses over all elements in the corresponding cluster. After fitting all the 2D anchor-poses into each of the candidate boxes, we perform class-specific regressions to deform these anchor-poses and match the actual 2D and 3D pose in each box. This

operation is carried out for the 3 types of parts, obtaining $5 \times J_B \times K_B$ outputs for the body part, $5 \times 2 \times J_H \times K_H$ for the hands and $5 \times J_F \times K_F$ for the face. The number 5 corresponds to the number of dimensions, *i.e.*, 2D+3D.

Postprocessing. For each body, hand or face, multiple proposals can overlap and produce valid predictions. As in [55], these pose candidates are combined, taking into account their 2D overlap, 3D similarity and classification scores. To obtain whole-body poses from the independent part detections produced by our network, we simply attach a hand to a body if their respective wrist estimations are close enough in 2D, and similarly for the face with the head body keypoint.

3.2 Distillation of part experts

Even if in-the-wild datasets with 3D pose annotations have been produced for bodies, hands and faces separately, there exists no dataset covering the whole-body at once. One possibility is to employ a union of these datasets to train our whole-body model. Since the datasets specifically designed for pose estimation of one part do not contain annotations for the others, *e.g.* body datasets do not have hand and face annotations and vice-versa, unannotated parts are therefore considered as negatives for their true classes in our detection architecture. In practice, this deteriorates the detector's ability to detect these parts and leads to worse overall performances ($\sim$10% drop for hands and faces, and $\sim$2% for bodies). To leverage the multiple part-specific datasets, we therefore propose to train independent experts for each part, namely body, hand and face experts, and distill their knowledge into our whole-body pose network designed to perform the three tasks jointly.

Part experts. To ease the distillation of the knowledge, we select our 3 experts to match the structure of the classification-regression branches of our whole-body pose estimation architecture and consider the same anchor poses as for the individual tasks. We therefore selected the Localization-Classification-Regression network from LCR-Net++ [55] as body expert and estimate $J_B = 13$ body joints with $K_B = 10$ classes. We also used the hand detection version of this architecture [2], replacing the K_B body pose classes by $K_H = 5$ hand anchor-poses for each side and using the standard number of $J_H = 21$ hand joints: 1 keypoint for the wrist plus 4 for each finger. Finally, to obtain our face expert, we adapted the same architecture to detect 2D-3D facial landmarks. We used the 84 landmarks defined in the 3D Face Tracking Menpo benchmark [69] that include eyes, eyebrows, nose, lips and facial contours. We defined $K_F = 10$ anchor-poses by applying K-means on all faces from the training set.

Training via distillation. We propose to distill the knowledge of our three part experts to our whole-body pose detection model. Let $\mathcal{B}$, $\mathcal{H}$ and $\mathcal{F}$ be the training datasets used for the three individuals tasks, *i.e.*, body, hand, and face pose detection, respectively. They are associated with ground-truth (2D and 3D) pose annotations for bodies b, hands h and faces f, respectively. In other words, the body expert is for instance trained on $\mathcal{B} = \{I_i, b_i\}_i$, *i.e.*, a set of images I_i with body ground-truth annotations b_i, and similarly for the other parts.

To train our network, we need ground-truth annotations w for the whole body. We propose to leverage the detections made by the experts in order to augment the annotations of the part-specific datasets. We denote by $\hat{b}_i$, $\hat{h}_i$ and $\hat{f}_i$ the detections obtained when processing the images I_i with our expert for body, hands and face respectively. We train our DOPE network on:

$$\mathcal{W}_{DOPE} = \{I_i, w_i\}_{i \in \mathcal{B} \cup \mathcal{H} \cup \mathcal{F}} \quad \text{where} \quad w_i = \begin{cases} \{b_i, \hat{h}_i, \hat{f}_i\} \text{ if } i \in \mathcal{B} \ , \\ \{\hat{b}_i, h_i, \hat{f}_i\} \text{ if } i \in \mathcal{H} \ , \\ \{\hat{b}_i, \hat{h}_i, f_i\} \text{ if } i \in \mathcal{F} \ . \end{cases} \tag{1}$$

The detections $\hat{b}_i$, $\hat{h}_i$ and $\hat{f}_i$ estimated by the experts are therefore considered as pseudo ground-truth for the missing keypoints in 2D and 3D. In practice, ground-truth annotations are completed using these estimations, for example when some annotations have been incorrectly labeled or are simply missing. Note that training images with no annotation at all could also be used to train our network, using only pseudo ground-truth annotations [39], *i.e.*, $w_i = \{\hat{b}_i, \hat{h}_i, \hat{f}_i\}$. The training scheme is illustrated in Figure 2.

Loss. Our loss $\mathcal{L}$ to train the network combines the RPN loss $\mathcal{L}_{RPN}$ as well as the sum of three terms for each part $p \in \{\text{body,hand,face}\}$: (a) a classification loss $\mathcal{L}^p_{cls}$, (b) a regression loss $\mathcal{L}^p_{reg}$, (c) a distillation loss $\mathcal{L}^p_{dist}$:

$$\mathcal{L} = \mathcal{L}_{RPN} + \sum_{p \in \{\text{body,hand,face}\}} \mathcal{L}^p_{cls} + \mathcal{L}^p_{reg} + \mathcal{L}^p_{dist} \ , \tag{2}$$

where $\mathcal{L}_{RPN}$ is the RPN loss from Faster R-CNN [53]. The classification loss $\mathcal{L}^p_{cls}$ for each part p is a standard softmax averaged over all boxes. If a box sufficiently overlaps with a ground-truth box, its ground-truth label is obtained by finding the closest anchor-pose from the ground-truth pose. Otherwise it is assigned a background label, *i.e.*, 0.

The regression loss $\mathcal{L}^p_{reg}$ is a standard L1 loss on the offset between ground-truth 2D-3D poses and their ground-truth anchor-poses, averaged over all boxes. Note that the regression is class-specific, and the loss is only applied on the output of the regressor specific to the ground-truth class for each positive box.

The distillation loss $\mathcal{L}^p_{dist}$ is composed of two elements, one for the distillation of the classification scores $\mathcal{L}^p_{dist\_cls}$ and another one, $\mathcal{L}^p_{dist\_reg}$, for the regression:

$$\mathcal{L}^p_{dist} = \mathcal{L}^p_{dist\_cls} + \mathcal{L}^p_{dist\_reg} \ . \tag{3}$$

Given a box, the goal of the distillation loss is to make the output of the whole-body network as close as possible to the output of the part expert p. The classification component $\mathcal{L}^p_{dist\_cls}$ is a standard distillation loss between the predictions produced by the corresponding part expert and those estimated by the whole-body model for part p. In other words, $\mathcal{L}^p_{dist\_cls}$ is the soft version of hard label loss $\mathcal{L}^p_{cls}$. The regression component $\mathcal{L}^p_{dist\_reg}$ is a L1 loss between the pose predicted by the part expert and the one estimated by the whole-body model for the ground-truth class. Note that the pseudo ground-truth pose is obtained by

averaging all overlapping estimates made by the part expert. While $\mathcal{L}^p_{reg}$ is designed to enforce regression of this pseudo ground-truth pose, $\mathcal{L}^p_{dist\_reg}$ favors regression of the exact same pose predicted by the part expert for a given box.

In practice, proposals generated by the RPNs of part experts and whole-body model are different but computing distillation losses requires some proposals to coincide. At training, we thus augment the proposals of the whole-body model with positive boxes from the part experts to compute these losses. In summary, given a training image, we: (a) run each part expert, keeping the positive boxes with classification scores and regression outputs, (b) run the whole-body model, adding the positive boxes from the experts to the list of proposals. Losses based on pseudo ground-truths are then averaged over all boxes while distillation losses are averaged only over positive boxes from the part experts.

3.3 Training details

Data. We train our body expert on the same combination of the MPII [1], COCO [41], LSP [35], LSPE [36], Human3.6M [33] and Surreal [62] datasets augmented with pseudo 3D ground-truth annotations as in [55]. We applied random horizontal flips while training for 50 epochs. We train our hand expert on the RenderedHand (RH) dataset [73] for 100 epochs, with color jittering, random horizontal flipping and perspective transforms. $K_H = 5$ anchor poses are obtained by clustering the 3D poses of right and flipped left hands from the training set. Finally, we train the face expert for 50 epochs on the Menpo dataset [69] with random horizontal flips and color jittering during training.
Implementation. We implement DOPE in Pytorch [50], following the Faster R-CNN implementation from Torchvision. We consider a ResNet50 backbone [28]. We train it for 50 epochs, using the union of the datasets of each part expert, simply doubling the RH dataset used for hands as the number of images is significantly lower than for the other parts. The same data augmentation strategy used for training each part expert is employed for the whole-body network. We use Stochastic Gradient Descent (SGD) with a momentum of 0.9, a weight decay of 0.0001 and an initial learning rate of 0.02, which is divided by 10 after 30 and 45 epochs. All images are resized such that the smallest image dimension is 800 pixels during training and testing and 1000 proposals are kept at test time.
Runtime. DOPE runs at 100ms on a single NVIDIA T4 GPU. When reducing the smallest image size to 400px and the number of box proposals to 50, and using half precision, it runs at 28 ms per image, *i.e.*, in real-time at 35 fps, with a 2-3% decrease of performance. For comparison, each of our experts runs at a similar framerate as our whole-body model since only the last layers change. Optimization-based 3D whole-body estimation methods [51, 64] take up to a minute to process each person.

4 Experiments

Given that there is no dataset to evaluate whole-body 3D pose estimation in the wild, we evaluate our method on each task separately. After presenting datasets

and metrics (Section 4.1), we compare the performance of our whole-body model to the experts (Section 4.2) and to the state of the art (Section 4.3).

4.1 Evaluation datasets and metrics

MPII for 2D body pose estimation. As in [55], we remove 1000 images from the MPII [1] training set and use them to evaluate our 2D body pose estimation results. We follow the standard evaluation protocol and report the PCKh@0.5 which is the percentage of correct keypoints with a keypoint being considered as correctly predicted if the error is smaller than half the size of the head.
MuPoTs for 3D body pose estimation. MuPoTs-3D [46] (Multi-person Pose estimation Test Set in 3D) is composed of more than 8,000 frames from 20 real-world scenes with up to three subjects. The ground-truth 3D poses, obtained using a multi-view MoCap system, have a slightly different format than the one estimated by our body expert and whole-body model. To better fit their 14-joint skeleton model, we modified the regression layer of our networks to output 14 keypoints instead of 13 while freezing the rest of the network. We finetuned this last layer only on the MuCo-3DHP dataset [46], the standard training set when testing on MuPoTs. We report the 3D-PCK, *i.e.*, the percentage of joint predictions with less than 15cm error, per sequence, and averaged over the subjects for which ground truth is available.
RenderedHand for 3D hand pose estimation. RenderedHand (RH) test set [73] consists of 2,728 images showing the hands of a single person. We report the standard AUC (Area Under the Curve) metric when plotting the 3D-PCK after normalizing the scale and relative translation between the ground-truth and the prediction. Note that while state-of-the-art methods evaluate hand pose estimation given ground-truth crops around the hands, we instead perform an automatic detection but miss around 2% of the hands.
Menpo for facial landmark estimation. We report results for facial landmark evaluation using the standard 3D-aware 2D metric [18] on the 30 videos from the test set of the ICCV'17 challenge [69]. Given a ground truth-matrix $\boldsymbol{s} \in \mathcal{M}_{N,2}(\mathbb{R})$ representing the 2D coordinates in the image of the $N = 84$ landmarks of a face, and a facial landmark prediction $\hat{\boldsymbol{s}} \in \mathcal{M}_{N,2}(\mathbb{R})$, this 2D normalized point-to-point RMS error is defined as:

$$\epsilon(\boldsymbol{s}, \hat{\boldsymbol{s}}) = \frac{\|\boldsymbol{s} - \hat{\boldsymbol{s}}\|_2}{\sqrt{N} d_{scale}} \ , \tag{4}$$

where d_{scale} is the length of the diagonal of the minimal 2D bounding box of $\boldsymbol{s}$.

4.2 Comparison to the experts

Table 1 presents a comparison of the performances obtained by the part experts and our DOPE model, for body, hand and face pose estimation tasks.

We first compare the part experts to a baseline where our whole-body network is trained on the partial ground-truth available for each dataset, *e.g.* only

Table 1: Comparison between our part experts and our whole-body model

	MPII (PCKh@0.5)	MuPoTs (PCK3D)	RH test (AUC)	Menpo (AUC)
body expert	**89.6**	66.8	-	-
hand expert	-	-	**87.1**	-
face expert	-	-	-	73.9
whole-body trained on gt	88.3	66.6	81.1	61.7
DOPE without $\mathcal{L}^p_{dist}$	88.3	66.4	83.5	**75.2**
DOPE with $\mathcal{L}^p_{dist}$	88.8	**67.2**	84.9	75.0

body annotations are available on images from body datasets, *etc*. The performance degrades quite significantly compared to those of the hand and face experts (-6% AUC for hand on the RH dataset and -12% for face landmarks). This is explained by a lower detection rate of the detector due to the fact that, for instance, unannotated faces present in the body datasets are considered as negatives during training. The performance of this model on body pose estimation is quite similar to the one of the body expert: as bodies are not observed too much in hand and face datasets, there are almost no missing body annotations.

We then compare the experts to a first version of our DOPE model without the distillation loss $\mathcal{L}^p_{dist}$. The performance on body pose estimation remains similar but, for hands and faces, a significant gain is obtained, in particular for faces, where the whole-body network performs even better than the expert. This might be explained by the fact that the whole-body network is trained on a larger variety of data, including images from body and hands datasets with many additional faces. In contrast, the hand component performs slightly lower than the expert. One hypothesis is that many hands in the body datasets are too small to be accurately estimated, leading to noisy pseudo ground-truth poses. However, the performance remains close to that of the hand expert.

With the addition of the distillation loss, the accuracy increases for hand pose estimation (+1.4%) and slightly for body pose estimation (+0.5% on MPII, +0.8% on MuPoTs), bringing the performance of the whole-body network even closer to the experts' results. Sometimes, DOPE even outperforms the part expert as observed on MuPoTs for multi-person 3D pose estimation or on Menpo for facial landmark detection. Figure 4 presents some qualitative results for the part experts and our proposed DOPE model trained with distillation loss. Two additional examples of our model's results can be found in Figure 1. DOPE produces high-quality whole-body detections that include bodies, hands and faces. In the example on the left in Figure 4, our whole-body network correctly detects and estimates the pose of three hands, misdetecting only the lady's right hand. By contrast, the hand expert only finds one hand in this image. Note that our method is holistic for each part: if a part is sufficiently visible in the image, a prediction is made for every keypoint of the part despite partial occlusions or truncations, as shown for the bodies in this same example. However, if a part is

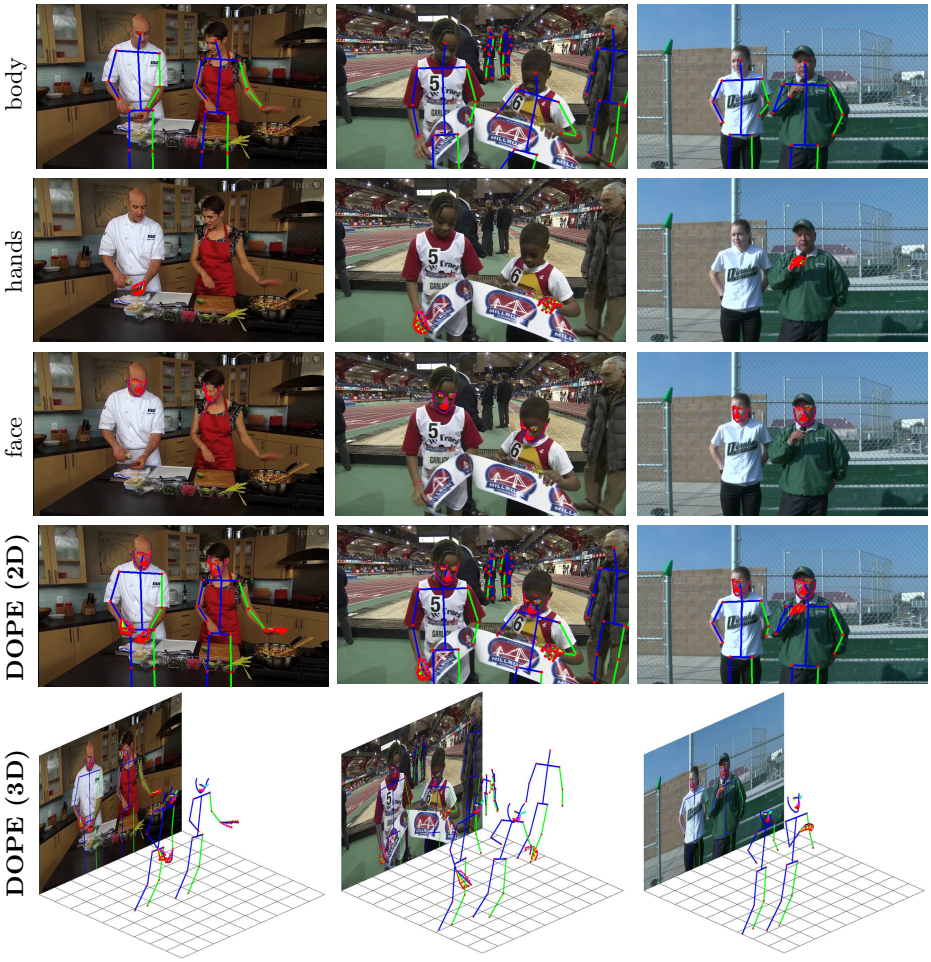
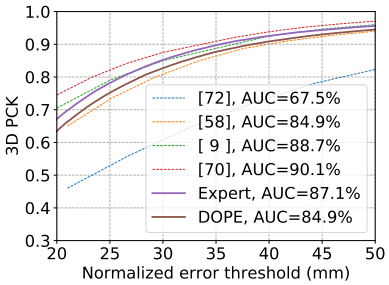


Fig. 4: Each column shows an example with the results of the 3 experts on the first three rows (we show only the 2D for clarity). The last two rows show the results obtained by our proposed DOPE approach in 2D and in 3D respectively.

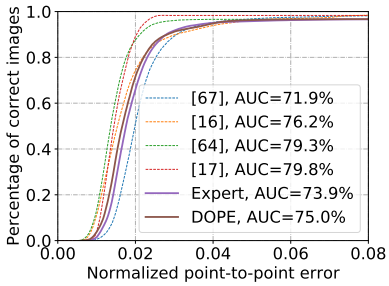
not visible, no prediction is made. This is the case for the occluded hands in the middle and right examples in Figure 4. Overall, these examples illustrate that our method can be applied in the wild, including scenes with multiple interacting people, varied background, severe occlusions or truncations.

4.3 Comparison to the state of the art

Comparison on individual tasks. In Figure 5, we compare our DOPE approach to the state of the art for each individual task. Note that our main goal is not to outperform the state of the art on each of these tasks but rather to unify 3 individual models into a single one while still achieving a competitive performance. DOPE is among the top performing methods for all three tasks,



(a)



(b)

method	PCK3D
VNect [47]	65.0
XNect [46]	70.4
LCRNet++ [55]	**70.6**
body expert	66.8
DOPE	67.2

(c)

method	RH (PCK2D)	Menpo (PCK2D)
Experts	**78.1**	94.5
DOPE	71.0	**94.9**
[12]	36.1	88.0
[29]	23.5	71.5

(d)

Fig. 5: Comparison to the state of the art: (a) PCK3D on RH for varying error threshold (hand). (b) Percentage of images with correct face detections for varying 3DA-2D thresholds on Menpo (face). (c) PCK3D on MuPoTs (body). (d) 2D PCK at a threshold of 10% of the tight bounding box's largest size on RH (hand) and 5% on Menpo (face). The higher the better.

i.e., hand (a), face (b) and body (c) 3D pose estimation, while being the first and only method to report on these three tasks together. Additionally, our detection network tackles a more difficult task than most of our competitors who assume that a bounding box around the ground-truth hands [10, 59, 71, 73] or faces [17, 18, 65, 68] is given at test time. We also compare with existing whole-body 2D pose estimation methods (d).

Qualitative comparison to [51, 64]. Since there is no dataset to numerically compare the performances of our learning-based approach in the wild against the optimization-based pipelines such as [51,64], we show some qualitative examples in Figure 6. We find that Monocular Total Capture [64] performs quite poorly on static images (second row), in particular due to occlusions. It greatly benefits from temporal information when processing the sequences from which the images are extracted (third row). However, there are still some errors, especially in case of occlusions (*e.g.* legs in the left column image). For [51] (fourth row), in the first example, OpenPose [12] does not estimate the 2D location of the feet that fall out of the field of view, impacting the optimization of the model's legs. In our case, the pose of the legs is correctly estimated. The same phenomenon happens in the second example where a little girl is kneeling and the self-occlusions prevent her feet from being detected. Finally, in the third example, the optimization gets stuck in a local minimum while our estimation is more robust. In addition of its robustness in the wild, our learning-based approach is also about 1000 times faster than [51] which takes about a minute per person in an image.

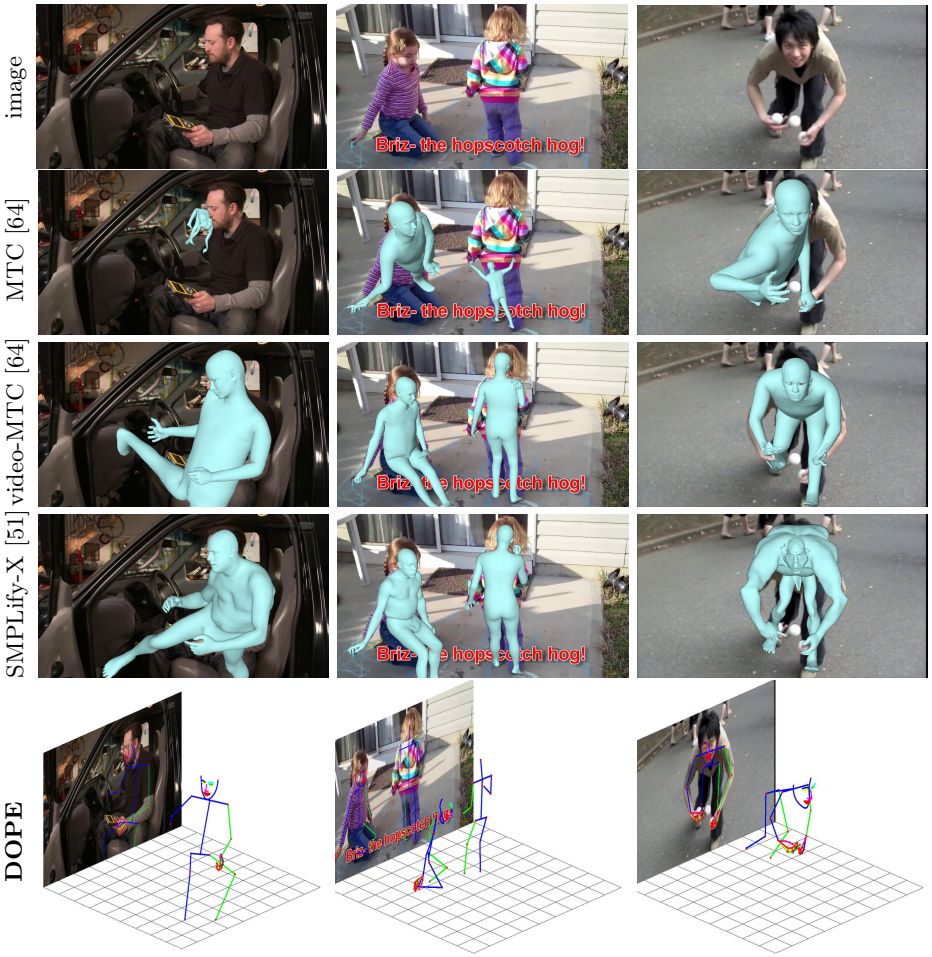


Fig. 6: Three examples with from top to bottom the original image, the results from MTC [64] from static image or the video, SMPLify-X [51] and ours.

5 Conclusion

We have proposed DOPE, the first learning-based method to detect and estimate whole-body 3D human poses in the wild, including body, hand and face 2D-3D keypoints. We tackled the lack of training data for this task by leveraging distillation from part experts to our whole-body network. Our experiments validated this approach showing performances close to the part experts' results.

Our method allows training a network on a more diverse set of in-the-wild images, potentially without any pose annotations. In future work, we will investigate if our model could benefit from additional unlabeled training data.

References

1. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2D human pose estimation: New benchmark and state of the art analysis. In: CVPR (2014)
2. Armagan, A., Garcia-Hernando, G., Baek, S., Hampali, S., Rad, M., Zhang, Z., Xie, S., Chen, M., Zhang, B., Xiong, F., Xiao, Y., Cao, Z., Yuan, J., Ren, P., Huang, W., Sun, H., Hrúz, M., Kanis, J., Krnoul, Z., Wan, Q., Li, S., Yang, L., Lee, D., Yao, A., Zhou, W., Mei, S., Liu, Y., Spurr, A., Iqbal, U., Molchanov, P., Weinzaepfel, P., Brégier, R., Rogez, G., Lepetit, V., Kim, T.: Measuring generalisation to unseen viewpoints, articulations, shapes and objects for 3D hand pose estimation under hand-object interaction. In: ECCV (2020)
3. Arnab, A., Doersch, C., Zisserman, A.: Exploiting temporal context for 3D human pose estimation in the wild. In: CVPR (2019)
4. Bhardwaj, S., Srinivasan, M., Khapra, M.M.: Efficient video classification using fewer frames. In: CVPR (2019)
5. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3D faces. In: SIGGRAPH (1999)
6. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: ECCV (2016)
7. Booth, J., Roussos, A., Zafeiriou, S., Ponniahy, A., Dunaway, D.: A 3D morphable model learnt from 10,000 faces. In: CVPR (2016)
8. Boukhayma, A., de Bem, R., Torr, P.H.S.: 3D hand shape and pose from images in the wild. In: CVPR (2019)
9. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In: ICCV (2017)
10. Cai, Y., Ge, L., Cai, J., Yuan, J.: Weakly-supervised 3D hand pose estimation from monocular RGB images. In: ECCV (2018)
11. Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K.: FaceWarehouse: a 3D facial expression database for visual computing. IEEE Transactions on Visualization and Computer Graphics (2013)
12. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In: arXiv preprint arXiv:1812.08008 (2018)
13. Chen, C.H., Ramanan, D.: 3D human pose estimation = 2D pose estimation + matching. In: CVPR (2017)
14. Chen, G., Choi, W., Yu, X., Han, T., Chandraker, M.: Learning efficient object detection models with knowledge distillation. In: NeurIPS (2017)
15. Cimen, G., Maurhofer, C., Sumner, B., Guay, M.: Ar poser: Automatically augmenting mobile pictures with digital avatars imitating poses. In: CGVCVIP (2018)
16. Crasto, N., Weinzaepfel, P., Alahari, K., Schmid, C.: MARS: Motion-augmented RGB stream for action recognition. In: CVPR (2019)
17. Crispell, D., Bazik, M.: Pix2Face: Direct 3D face model estimation. In: ICCV Workshop (2017)
18. Deng, J., Roussos, A., Chrysos, G., Ververas, E., Kotsia, I., Shen, J., Zafeiriou, S.: The Menpo benchmark for multi-pose 2D and 3D facial landmark localisation and tracking. IJCV (2019)
19. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: RMPE: Regional multi-person pose estimation. In: ICCV (2017)

20. Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X.: Joint 3D face reconstruction and dense alignment with position map regression network. In: ECCV (2020)
21. Garcia-Salguero, M., Gonzalez-Jimenez, J., Moreno, F.A.: Human 3D pose estimation with a tilting camera for social mobile robot interaction. Sensors (2019)
22. Ge, L., Ren, Z., Li, Y., Xue, Z., Wang, Y., Cai, J., Yuan, J.: 3D hand shape and pose estimation from a single RGB image. In: CVPR (2019)
23. Gui, L.Y., Zhang, K., Wang, Y.X., Liang, X., Moura, J.M., Veloso, M.: Teaching robots to predict human motion. In: IROS (2018)
24. Habibie, I., Xu, W., Mehta, D., Pons-Moll, G., Theobalt, C.: In the wild human pose estimation using explicit 2D features and intermediate 3D representations. In: CVPR (2019)
25. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: Honnotate: A method for 3D annotation of hand and objects poses. In: CVPR (2020)
26. Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M.J., Laptev, I., Schmid, C.: Learning joint reconstruction of hands and manipulated objects. In: CVPR (2019)
27. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV (2017)
28. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
29. Hidalgo, G., Raaj, Y., Idrees, H., Xiang, D., Joo, H., Simon, T., Sheikh, Y.: Single-network whole-body pose estimation. In: ICCV (2019)
30. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NIPS workshop (2014)
31. Hoffman, J., Gupta, S., Darrell, T.: Learning with side information through modality hallucination. In: CVPR (2016)
32. Hou, S., Pan, X., Change Loy, C., Wang, Z., Lin, D.: Lifelong learning via progressive distillation and retrospection. In: ECCV (2018)
33. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE trans. PAMI (2013)
34. Jackson, A.S., Bulat, A., Argyriou, V., Tzimiropoulos, G.: Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In: ICCV (2017)
35. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: BMVC (2010)
36. Johnson, S., Everingham, M.: Learning effective human pose estimation from inaccurate annotation. In: CVPR (2011)
37. Joo, H., Simon, T., Sheikh, Y.: Total capture: A 3D deformation model for tracking faces, hands, and bodies. In: CVPR (2018)
38. Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M.J., Gehler, P.V.: Unite the people: Closing the loop between 3D and 2D human representations. In: CVPR (2017)
39. Lee, D.H.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: ICML Workshop (2013)
40. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4D scans. ACM Transactions on Graphics (ToG) (2017)
41. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
42. Liu, X., He, P., Chen, W., Gao, J.: Improving multi-task deep neural networks via knowledge distillation for natural language understanding. arXiv preprint arXiv:1904.09482 (2019)

43. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM transactions on Graphics (2015)
44. Lopez-Paz, D., Bottou, L., Schölkopf, B., Vapnik, V.: Unifying distillation and privileged information. In: ICLR (2016)
45. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3D human pose estimation. In: ICCV (2017)
46. Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., Theobalt, C.: Single-shot multi-person 3D pose estimation from monocular RGB. In: 3DV (2018)
47. Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.P., Xu, W., Casas, D., Theobalt, C.: VNect: Real-time 3D human pose estimation with a single RGB camera. ACM Transactions on Graphics (2017)
48. Moon, G., Chang, J.Y., Lee, K.M.: Camera distance-aware top-down approach for 3D multi-person pose estimation from a single RGB image. In: ICCV (2019)
49. Mueller, F., Bernard, F., Sotnychenko, O., Mehta, D., Sridhar, S., Casas, D., Theobalt, C.: GANerated hands for real-time 3D hand tracking from monocular RGB. In: CVPR (2018)
50. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: NeurIPS (2019)
51. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: CVPR (2019)
52. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3D human pose. In: CVPR (2017)
53. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: NIPS (2015)
54. Rogez, G., Schmid, C.: Mocap-guided data augmentation for 3D pose estimation in the wild. In: NIPS (2016)
55. Rogez, G., Weinzaepfel, P., Schmid, C.: LCR-Net++: Multi-person 2D and 3D pose detection in natural images. IEEE trans. PAMI (2019)
56. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: modeling and capturing hands and bodies together. ACM Transactions on Graphics (2017)
57. Sanyal, S., Bolkart, T., Feng, H., Black, M.J.: Learning to regress 3D face shape and expression from an image without 3D supervision. In: CVPR (2019)
58. Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: CVPR (2017)
59. Spurr, A., Song, J., Park, S., Hilliges, O.: Cross-modal deep variational hand pose estimation. In: CVPR (2018)
60. Supani, J.S., Rogez, G., Yang, Y., Shotton, J., Ramanan, D.: Depth-based hand pose estimation: Methods, data, and challenges. IJCV (2018)
61. Vapnik, V., Izmailov, R.: Learning using privileged information: Similarity control and knowledge transfer. JMLR (2015)
62. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: CVPR (2017)
63. Wu, Y., Ji, Q.: Facial landmark detection: a literature survey. IJCV (2019)
64. Xiang, D., Joo, H., Sheikh, Y.: Monocular total capture: Posing face, body, and hands in the wild. In: CVPR (2019)

65. Xiong, P., Li, G., Sun, Y.: Combining local and global features for 3D face tracking. In: ICCV Workshops (2017)
66. Yang, L., Li, S., Lee, D., Yao, A.: Aligning latent spaces for 3D hand pose estimation. In: ICCV (2019)
67. Yuan, S., Stenger, B., Kim, T.K.: RGB-based 3D hand pose estimation via privileged learning with depth images. arXiv preprint arXiv:1811.07376 (2018)
68. Zadeh, A., Baltrusaitis, T., Morency, L.P.: Convolutional experts constrained local model for facial landmark detection. In: CVPR Workshop (2017)
69. Zafeiriou, S., Chrysos, G., Roussos, A., Ververas, E., Deng, J., Trigeorgis, G.: The 3D menpo facial landmark tracking challenge. In: ICCV Workshops (2017)
70. Zhang, J., Jiao, J., Chen, M., Qu, L., Xu, X., Yang, Q.: A hand pose tracking benchmark from stereo matching. In: ICIP (2017)
71. Zhang, X., Li, Q., Mo, H., Zhang, W., Zheng, W.: End-to-end hand mesh recovery from a monocular RGB image. In: ICCV (2019)
72. Zhu, X., Liu, X., Lei, Z., Li, S.Z.: Face alignment in full pose range: A 3D total solution. IEEE trans. PAMI (2017)
73. Zimmermann, C., Brox, T.: Learning to estimate 3D hand pose from single RGB images. In: ICCV (2017)