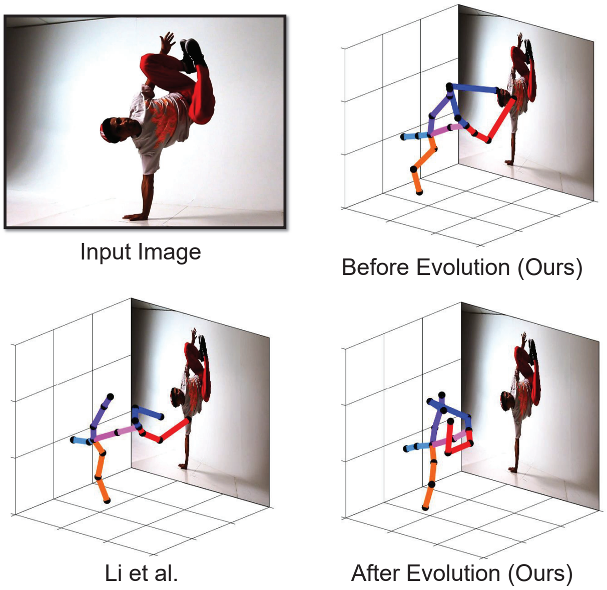
Cascaded Deep Monocular 3D Human Pose Estimation with Evolutionary Training Data

Shichao Li[1], Lei Ke[1], Kevin Pratama[1], Yu-Wing Tai[2], Chi-Keung Tang[1], Kwang-Ting Cheng[1]
[1]The Hong Kong University of Science and Technology, [2]Tencent

Abstract

End-to-end deep representation learning has achieved remarkable accuracy for monocular 3D human pose estimation, yet these models may fail for unseen poses with limited and fixed training data. This paper proposes a novel data augmentation method that: (1) is scalable for synthesizing massive amount of training data (over 8 million valid 3D human poses with corresponding 2D projections) for training 2D-to-3D networks, (2) can effectively reduce dataset bias. Our method evolves a limited dataset to synthesize unseen 3D human skeletons based on a hierarchical human representation and heuristics inspired by prior knowledge. Extensive experiments show that our approach not only achieves state-of-the-art accuracy on the largest public benchmark, but also generalizes significantly better to unseen and rare poses. Relevant files and tools are available at the project website[1][2].



Input Image

Before Evolution (Ours)

Li et al.

After Evolution (Ours)

Figure 1: Model trained on the evolved training data generalizes better than [25] to unseen inputs.

1. Introduction

Estimating 3D human pose from RGB images is critical for applications such as action recognition [32] and human-computer interaction, yet it is challenging due to lack of depth information and large variation in human poses, camera viewpoints and appearances. Since the introduction of large-scale motion capture (MC) datasets [56, 20], learning-based methods and especially deep representation learning have gained increasing momentum in 3D pose estimation. Thanks to their representation learning power, deep models have achieved unprecedented high accuracy [43, 41, 28, 34, 32, 60].

Despite their success, deep models are data-hungry and vulnerable to the limitation of data collection. This problem is more severe for 3D pose estimation due to two factors. First, collecting accurate 3D pose annotation for RGB images is expensive and time-consuming. Second, the collected training data is usually biased towards indoor environment and selected daily actions. Deep models can easily exploit these bias but fail for unseen cases in unconstrained environments. This fact has been validated by recent works [69, 66, 25, 64] where cross-dataset inference demonstrated poor generalization of models trained with biased data.

To cope with the domain shift of appearance for 3D pose estimation, recent state-of-the-art (SOTA) deep models adopt the two-stage architecture [68, 13, 14]. The first stage locates 2D human key-points from appearance information, while the second stage lifts the 2D joints into 3D skeleton employing geometric information. Since 2D pose annotations are easier to obtain, extra in-the-wild images can be used to train the first stage model, which effectively reduces the bias towards indoor images during data collection. However, the second stage 2D-to-3D model can still be negatively influenced by geometric data bias, yet not studied before. We focus on this problem in this work and our research questions are: *are our 2D-to-3D deep networks influenced by data bias? If yes, how can we improve network*

[1] https://github.com/Nicholasli1995/EvoSkeleton
[2] The arxiv version will present future updates if any.

generalization when the training data is limited in scale or variation?

To answer these questions, we propose to analyze the training data with a hierarchical human model and represent human posture as collection of local bone orientations. We then propose a novel dataset evolution framework to cope with the limitation of training data. Without any extra annotation, we define evolutionary operators such as *crossover* and *mutation* to discover novel valid 3D skeletons in tree-structured data space guided by simple prior knowledge. These synthetic skeletons are projected to 2D and form 2D-3D pairs to augment the data used for training 2D-to-3D networks. With an augmented training dataset after evolution, we propose a cascaded model achieving state-of-the-art accuracy under various evaluation settings. Finally, we release a new dataset for unconstrained humans in-the-wild. Our contributions are summarized as follows:

- To our best knowledge, we are the first to improve 2D-to-3D network training with synthetic paired supervision.
- We propose a novel data evolution strategy which can augments an existing dataset by exploring 3D human pose space without intensive collection of extra data. This approach is scalable to produce 2D-3D pairs in the order of 10^7, leading to better model generalization power in unseen scenarios.
- We present TAG-Net, a deep architecture consisting of an accurate 2D joint detector and a novel cascaded 2D-to-3D network. It out-performs previous monocular models on the largest 3D human pose estimation benchmark in various aspects.
- We release a new labeled dataset for unconstrained human pose estimation in-the-wild.

Fig. 1 shows our model trained on an augmented dataset can handle rare poses while others such as [25] may fail.

2. Related Works

Monocular 3D human pose estimation. Single-image 3D pose estimation methods are conventionally categorized into *generative methods* and *discriminative methods*. Generative methods fit parametrized models to image observations for 3D pose estimation. These approaches represent humans by PCA models [2, 70], graphical models [8, 5] or deformable meshes [4, 30, 7, 42, 24]. The fitting process amounts to non-linear optimization, which requires good initialization and refines the solution iteratively. Discriminative methods [53, 1, 6] directly learn a mapping from image observations to 3D poses. Pertinent and recent deep neural networks (DNNs) employ two mainstream architectures: *one-stage* methods [66, 69, 32, 43, 41, 28, 60, 16] and *two-stage* methods [39, 34, 47, 68]. The former directly map from pixel intensities to 3D poses, while the latter first extract intermediate geometric representation such as 2D key-points and then lift them to 3D poses.

We adopt the discriminative approach and focus on the second stage. Instead of using a fixed training dataset, we evolve the training data to improve the performance of the 2D-to-3D network.

Weakly-supervised 3D pose estimation. Supervised training of DNNs demands massive data while 3D annotation is difficult. To address this problem, weakly-supervised methods explore other potential supervision to improve network performance when only few training data is available [44, 49, 50, 23, 11, 65, 27]. Multi-view consistency [44, 49, 50, 23, 11] is proposed and validated as useful supervisory signal when training data is scarce, yet a minimum of two views are needed. In contrast, we focus on effective utilization of scarce training data by synthesizing new data from existing ones and *uses only single view*.

Data augmentation for pose estimation. New images can be synthesized to augment indoor training dataset [51, 63]. In [63] new images were rendered using MC data and human models. Domain adaption was performed in [10] during training with synthetic images. Adversarial rotation and scaling were used in [46] to augment data for 2D pose estimation. These works produce augmented images while we focus on data augmentation for 2D-to-3D networks and produce geometric 2D-3D pairs.

Pose estimation dataset. Most large-scale human pose estimation datasets [67, 29, 3] only provide 2D pose annotations. Accurate 3D annotations [20, 56] require MC devices and these datasets are biased due to the limitation of data collection process. Deep models are prone to overfit to these biased dataset [61, 62, 26], failing to generalize in unseen situations. Our method can synthesize *for free without human annotation* large amount of valid 3D poses with more complete coverage in human pose space.

3. Dataset Evolution

From a given input image $\mathbf{x}_i$ containing one human subject, we aim to infer the 3D human pose $\hat{\mathbf{p}}_i$ given the image observation $\phi(\mathbf{x}_i)$. To encode geometric information as other 2D-to-3D approaches [34, 68, 25], we represent $\phi(\mathbf{x})$ as the 2D coordinates of k human key-points $(x_i, y_i)_{i=1}^{k}$ on the image plane. As a discriminative approach, we seek a regression function $\mathcal{F}(\phi(\mathbf{x}_i), \mathbf{\Theta})$ that outputs 3D pose $\hat{\mathbf{p}}_i$ as $\hat{\mathbf{p}}_i = \mathcal{F}(\phi(\mathbf{x}_i), \mathbf{\Theta})$. This regression function is implemented as a DNN parametrized by $\mathbf{\Theta}$. Conventionally this DNN is trained on a dataset collected by MC devices [56, 20]. This dataset consists of paired images and 3D pose ground truths $\{(\mathbf{x}_i, \mathbf{p}_i)\}_{i=1}^{N}$ and the DNN can be trained by gradient descent based on a loss function defined over the training dataset $\mathcal{L} = \sum_{i=1}^{N} E(\mathbf{p}_i, \mathcal{F}(\phi(\mathbf{x}_i), \mathbf{\Theta}))$

where E is the error measurement between the ground truth $\mathbf{p}_i$ and the prediction $\hat{\mathbf{p}}_i = \mathcal{F}(\phi(\mathbf{x}_i), \mathbf{\Theta})$.

Unfortunately, sampling bias exists during the data collection and limits the variation of the training data. Human 3.6M (H36M) [20], the largest MC dataset, only contains 11 subjects performing 15 actions under 4 viewpoints, leading to insufficient coverage of the training 2D-3D pairs $(\phi(\mathbf{x}_i), \mathbf{p}_i)$. A DNN can overfit to the dataset bias and become less robust to unseen $\phi(\mathbf{x})$. For example, when a subject starts street dancing, the DNN may fail since it is only trained on daily activities such as sitting and walking. This problem is even exacerbated for the weakly-supervised methods [44, 50, 11] where only a subset of training data is used to simulate the data scarcity scenario.

We take a non-stationary view toward the training data to cope with this problem. While conventionally the collected training data is fixed and the trained DNN is not modified during its deployment, here we assume the data and model can evolve during their life-time. Specifically, we synthesize novel 2D-3D pairs based on an initial training dataset and add them into the original dataset to form the evolved dataset. We then re-train the model with the evolved dataset. As shown in Fig. 2, model re-trained on the evolved dataset has consistently lower generalization error, comparing to a model trained on the initial dataset.

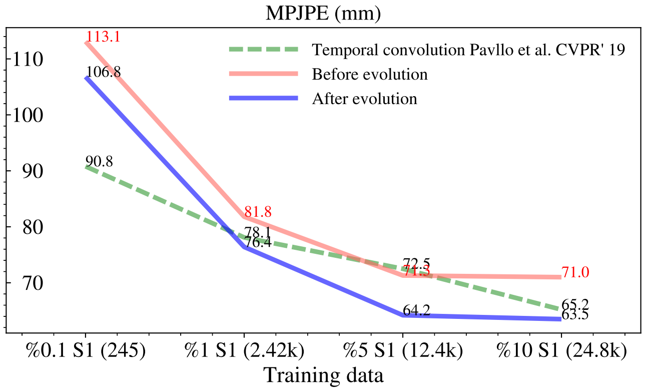


Figure 2: Generalizing errors (MPJPE using ground truth 2D key-points as inputs) on H36M before and after dataset evolution with varying size of initial population.

In the following we show that by using a hierarchical representation of human skeleton, the synthesis of novel 2D-3D pairs can be achieved by evolutionary operators and camera projection.

3.1. Hierarchical Human Representation

We represent a 3D human skeleton by a set of bones organized hierarchically in a kinematic tree as shown in Figure 3. This representation captures the dependence of adjacent joints using tree edges.

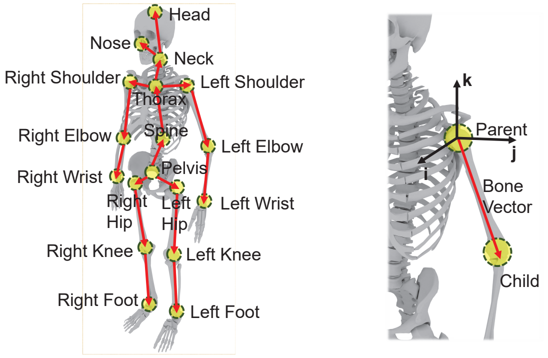


Figure 3: Hierarchical human representation. Left: 3D key-points organized in a kinematic tree where red arrows point from parent joints to children joints. Right: Zoom-in view of a local coordinate system.

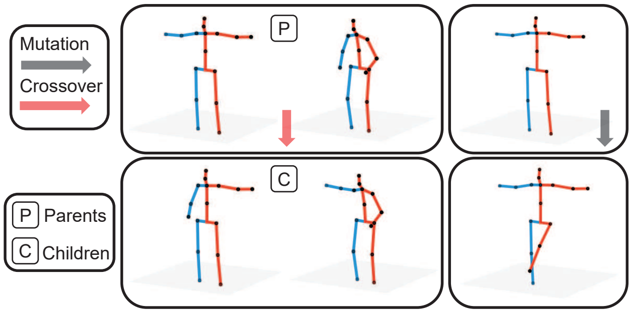


Figure 4: Examples of applying evolution operators. Crossover and mutation take 2 and 1 random samples respectively to synthesize novel human skeletons.

Each 3D pose $\mathbf{p}$ corresponds to a set of bone vectors $\{\mathbf{b}^1, \mathbf{b}^2, \cdots, \mathbf{b}^w\}$ and a bone vector is defined as

$$\mathbf{b}^i = \mathbf{p}^{child(i)} - \mathbf{p}^{parent(i)} \tag{1}$$

where $\mathbf{p}^j$ is the jth joint in the 3D skeleton and $parent(i)$ gives the parent joint index of the ith bone vector. A local coordinate system[3] is attached at each parent node. For a parent node $\mathbf{p}^{parent(i)}$, its local coordinate system is represented by the rotation matrix defined by three basis vectors $\mathbf{R}^i = [\mathbf{i}^i, \mathbf{j}^i, \mathbf{k}^i]$. The global bone vector is transformed into this local coordinate system as

$$\mathbf{b}^i_{local} = \mathbf{R}^{i\mathbf{T}} \mathbf{b}^i_{global} = \mathbf{R}^{i\mathbf{T}} (\mathbf{p}^{child(i)} - \mathbf{p}^{parent(i)}) \tag{2}$$

For convenience, this local bone vector is further converted into spherical coordinates $\mathbf{b}^i_{local} = (r_i, \theta_i, \phi_i)$. The posture of the skeleton can be described by the collection of bone orientations $\{(\theta_i, \phi_i)\}_{i=1}^{w}$ while the skeleton size is encoded into $\{r_i\}_{i=1}^{w}$.

[3] The coordinate system is detailed in our supplementary material.

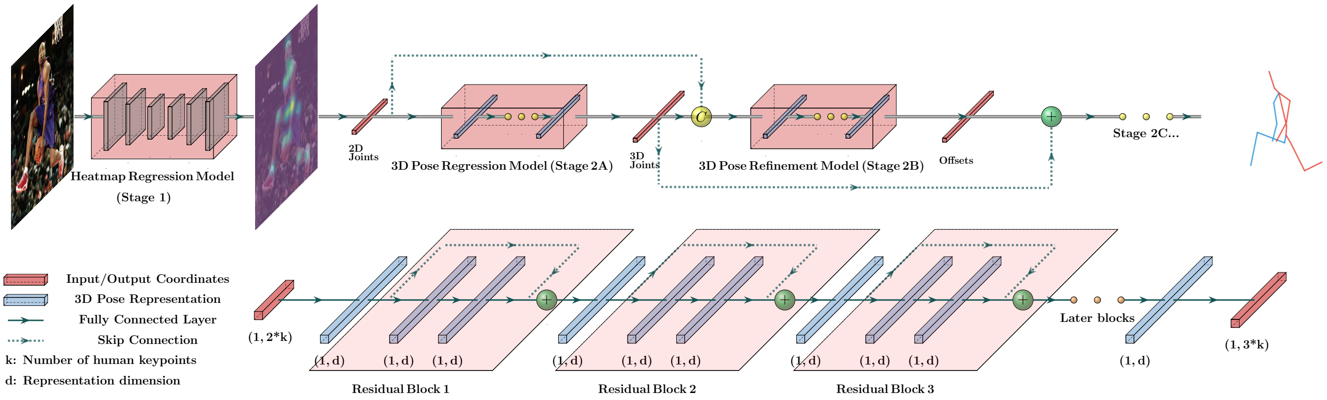


Figure 5: Our cascaded 3D pose estimation architecture. Top: our model is a two-stage model where the first stage is a 2D landmark detector and the second stage is a cascaded 3D coordinate regression model. Bottom: each learner in the cascade is a feed-forward neural network whose capacity can be adjusted by the number of residual blocks. To fit an evolved dataset, we use 8 layers (3 blocks) for each cascade and have 24 layers in total with a cascade of 3 models.

3.2. Synthesizing New 2D-3D Pairs

We first synthesize new 3D skeletons $\mathcal{D}_{new} = \{\mathbf{p}_j\}_{j=1}^{M}$ with an initial training dataset $\mathcal{D}_{old} = \{\mathbf{p}_i\}_{i=1}^{N}$ and project 3D skeletons to 2D given camera intrinsics $\mathbf{K}$ to form 2D-3D pairs $(\phi(\mathbf{x}_j), \mathbf{p}_j)$ where $\phi(\mathbf{x}_j) = \mathbf{K}\mathbf{p}_j$.

When adopting the hierarchical representation, a dataset of articulated 3D objects is a population of tree-structured data in nature. Evolutionary operators [18] have *constructive* property [57] that can be used to synthesize new data [15] given an initial population. The design of operators is problem-dependent and our operators are detailed as follows.

Crossover Operator Given two parent 3D skeletons, crossover is defined as a random exchange of sub-trees. This definition is inspired by the observation that an unseen 3D pose might be obtained by assembling limbs from known poses. Formally, we denote the set of bone vectors for parent A and B as $S_A = \{\mathbf{b}_A^1, \mathbf{b}_A^2, \ldots, \mathbf{b}_A^w\}$ and $S_B = \{\mathbf{b}_B^1, \mathbf{b}_B^2, \ldots, \mathbf{b}_B^w\}$. A joint indexed by q is selected at random and the bones rooted at it are located for the two parents. These bones form the chosen sub-tree set S_{chosen}

$$\{\mathbf{b}^j : parent(j) = q \vee \mathit{IsOff}(parent(j), q)\} \tag{3}$$

where $\mathit{IsOff}(parent(j), q)$ is $True$ if joint $parent(j)$ is an offspring of joint q in the kinematic tree. The parent bones are split into the chosen and the remaining ones as $S_X = S_{chosen}^X \cup S_{rem}^X$where $S_{rem}^X = S_X - S_{chosen}^X$ and X is A or B. Now the crossover operator gives two sets of children bones as

$$S_C = S_{chosen}^A \cup S_{rem}^B \quad \text{and} \quad S_D = S_{chosen}^B \cup S_{rem}^A \tag{4}$$

These two new sets are converted into two new 3D skeletons. The example in Fig. 4 shows the exchange of the right arms when the right shoulder joint is selected.

Algorithm 1 Data evolution

```
Input:
Initial set of 3D skeletons $D_{old} = \{\mathbf{p}_i\}_{i=1}^{N}$, noise level $\sigma$, number of generations $G$
Output: Augmented set of skeletons $D_{new} = \{\mathbf{p}_i\}_{i=1}^{M}$
$D_{new} = D_{old}$
for i=1:$G$ do
    Parents = Sample($D_{new}$)
    Children = NaturalSelection(Mutation(Crossover(Parents)))
    $D_{new} = D_{new} \cup$ Children
end for
return $D_{new}$
```

Mutation Operator As the motion of human limbs is usually continuous, a perturbation of one limb of an old 3D skeleton may result in a valid new 3D pose. To implement this perturbation, our mutation operator modifies the local orientation of one bone vector to get a new pose. One bone vector $\mathbf{b}_i = (r_i, \theta_i, \phi_i)$ for an input 3D pose is selected at random and its orientation is mutated by adding noise (Gaussian in this study):

$$\theta_i' = \theta_i + g, \phi_i' = \phi_i + g \tag{5}$$

where $g \sim N(0, \sigma_{local})$ and σ_{local} is a pre-defined noise level. One example of mutating the left leg is shown in Fig. 4. We also mutate the global orientation and bone length of the 3D skeletons to reduce the data bias of viewpoints and subject sizes, which is detailed in our supplementary material.

Natural Selection We use a fitness function to evaluate the goodness of synthesized data for selection as $v(\mathbf{p})$ which indicates the validity of the new pose. $v(\mathbf{p})$ can be any function that describes how anatomically valid a skeleton is, and we implement it by utilizing the binary function provided by [2]. We specify $v(\mathbf{p}) = -\infty$ if $\mathbf{p}$ is not valid to rule out all invalid poses.

Evolution Process The above operators are applied to D_{old} to obtain a new generation D_{new} by synthesizing new poses and merge with the old poses. This evolution process repeats several generations and is depicted in Algorithm 1. Finally, D_{new} are projected to 2D key-points to obtain paired 2D-3D supervision.

4. Model Architecture

We propose a two-stage model as shown in Fig. 5. We name it TAG-Net, as the model's focus **t**ransits from **a**ppearance to **g**eometry. This model can be represented as a function

$$\hat{\mathbf{p}} = TAG(\mathbf{x}) = \mathcal{G}(\mathcal{A}(\mathbf{x})) \tag{6}$$

Given an input RGB image $\mathbf{x}$, $\mathcal{A}(\mathbf{x})$ (the appearance stage) regresses $k = 17$ high-resolution probability heat-maps $\mathbf{H}_{i=1}^{k}$ for k 2D human key-points and map them into 2D coordinates $\mathbf{c} = (x_i, y_i)_{i=1}^{k}$. $\mathcal{G}(\mathbf{c})$ (the geometry stage) infers 3D key-point coordinates[4] $\mathbf{p} = (x_i, y_i, z_i)_{i=1}^{k}$ in the camera coordinate system from input 2D coordinates. Key designs are detailed as follows.

4.1. High-resolution Heatmap Regression

Synthesized 2D key-points are projected from 3D points and can be thought as perfect detections while real detections produced by heat-map regression models are noisier. We hope this noise can be as small as possible since we need to merge these two types of data as described in Section 3. To achieve this goal, we use HR-Net [59] as our backbone for image feature extraction. While the original model predicts heat-maps of size 96 by 72, we append a pixel shuffle layer [55] to the end and regress heat-maps of size 384 by 288. The original model uses hard arg-max to predict 2D coordinates, which results in rounding errors in our experiments. Instead, we use soft arg-max [40, 60] to obtain 2D coordinates. The average 2D key-point localization errors for H36M testing images are shown in Table 1. Our design choice improves the previous best model and achieves the *highest* key-point localization accuracy on H36M to date. The extensions add negligible amount of parameters and computation.

Backbone	Extension	#Params	FLOPs	Error
CPN [12]	-	-	13.9G	5.40
HRN [59]	-	63.6M	32.9G	$4.98_{\downarrow 7.8\%}$
HRN	+ U	63.6M	32.9G	$4.64_{\downarrow 14.1\%}$
HRN	+ U + S	63.6M	32.9G	$4.36_{\downarrow 19.2\%}$

Table 1: Average 2D key-point localization errors for H36M testing set in terms of pixels. U: Heat-map up-sampling. S: use soft-argmax. Error reduction compared with the previous best model [12] used in [45] follows the ↓ signs.

4.2. Cascaded Deep 3D Coordinate Regression

Since the mapping from 2D coordinates to 3D joints can be highly-nonlinear and difficult to learn, we propose a cascaded 3D coordinate regression model as

$$\hat{\mathbf{p}} = \mathcal{G}(\mathbf{c}) = \sum_{t=1}^{T} \mathcal{D}_t(\mathbf{i}_t, \mathbf{\Theta}_t) \tag{7}$$

where $\mathcal{D}_t$ is the tth *deep* learner in the cascade parametrized by $\mathbf{\Theta}_t$ whose input is $\mathbf{i}_t$. As shown in the top of Fig. 5, the first learner $\mathcal{D}_1$ in the cascade directly predicts 3D pose while the later ones predict the 3D refinement $\delta\mathbf{p} = (\delta x_i, \delta y_i, \delta z_i)_{i=1}^{k}$. While cascaded coordinate regression has been adopted for 2D key-points localization [9, 48], hand-crafted image feature and classical weak learners such as linear regressors were used. In contrast, our geometric model $\mathcal{G}(\mathbf{c})$ only uses coordinates as input and each learner is a DNN with residual connections [17].

The bottom of Fig. 5 shows the detail for each deep learner. One deep learner first maps the input 2D coordinates into a representation vector of dimension $d = 1024$, after which $R = 3$ residual blocks are used. Finally the representation is mapped by a fully-connected (FC) layer into 3D coordinates. After each FC layer we add batch normalization [19] and dropout [58] with dropout rate 0.5. The capacity of each deep learner can be controlled by R. This cascaded model is trained sequentially by gradient descent and the training algorithm is included in our supplementary material. Despite the number of parameters increase linearly with the cascade length, we found that the cascaded model is robust to over-fitting for this 3D coordinate prediction problem, which is also shared by the 2D counterparts [9, 48].

4.3. Implementation Details

We train $\mathcal{A}(\mathbf{x})$ and $\mathcal{G}(\mathbf{c})$ sequentially. The input size is 384 by 288 and our output heat-map has the same high resolution. The back-bone of $\mathcal{A}(\mathbf{x})$ is pre-trained on COCO [29] and we fine-tune it on H36M with Adam optimizer using a batch size of 24. The training is performed on two NVIDIA Titan Xp GPUs and takes 8 hours for 18k iterations. We first train with learning rate 0.001 for 3k iterations, after which we multiply it by 0.1 after every 3k iterations. To train $\mathcal{G}(\mathbf{c})$, we train each deep learner in the cascade using Adam optimizer with learning rate 0.001 for 200 epochs.

5. Experiments

To validate our data evolution framework and model architecture, we evolve from the training data provided in

[4] Relative to the root joint.

H36M and conduct both intra- and cross-dataset evaluation. The camera intrinsics provided by H36M are used during data synthesis. We vary the size of initial population to demonstrate the effectiveness of synthetic data when the training data is scarce. Finally we present ablation study to analyze the influences of data augmentation and hyper-parameters.

5.1. Datasets and Evaluation Metrics

Human 3.6M (H36M) is the largest 3D human pose estimation benchmark with accurate 3D labels. We denote a collection of data by appending subject ID to S, e.g., S15 denotes data from subject 1 and 5. Previous works fix the training data while our method uses it as our initial population and evolves from it. We evaluate model performance with *Mean Per Joint Position Error* (MPJPE) measured in millimeters. Two standard evaluation protocols are adopted. *Protocol 1* (P1) directly computes MPJPE while *Protocol 2* (P2) aligns the ground-truth 3D poses with the predictions with a rigid transformation before calculating it. Protocol P1* uses ground truth 2D key-points as inputs and removes the influence of the first stage model.

MPI-INF-3DHP (3DHP) is a benchmark that we use to evaluate the generalization power of 2D-to-3D networks. We do not use its training data and conduct cross-dataset inference by feeding the provided key-points to $\mathcal{G}(\mathbf{c})$. Apart from MPJPE, *Percentage of Correct Keypoints* (PCK) measures correctness of 3D joint predictions under a specified threshold, while *Area Under the Curve* (AUC) is computed for a range of PCK thresholds.

Unconstrained 3D Poses in the Wild (U3DPW) We collect by ourselves a new small dataset consisting of 300 challenging in-the-wild images with rare human poses, where 150 of them are selected from Leeds Sports Pose dataset [21]. The annotation process is detailed in our supplementary material. Similar to 3DHP, this dataset is used for validating model generalization for unseen 3D poses.

5.2. Comparison with state-of-the-art methods

Comparison with weakly-supervised methods Here we compare with *weakly-supervised methods*, which only use a small number of training data to simulate scarce data scenario. To be consistent with others, we utilize S1 as our initial population. While others fix S1 as the training dataset, we evolve from it to obtain an augmented training set. The comparison of model performance is shown in Table 2, where our model significantly out-performs others and demonstrates effective use of the limited training data. While other methods [50, 23] use multi-view consistency as extra supervision, we achieve comparable performance with only a single view by synthesizing useful supervision. Fig. 2 validates our method when the training data is extremely scarce, where we start with a small fraction of S1 and increase the data size by 2.5 times by evolution. Note that the model performs consistently better after dataset evolution. Compared to the temporal convolution model proposed in [45], we do not utilize any temporal information and achieve comparable performance. This indicates our approach can make better use of extremely limited data.

Method	**Performance**		
Authors	P1	P1*	P2
Use Multi-view			
Rhodin *et al*. (CVPR'18) [50]	-	-	64.6
Kocabas *et al*. (CVPR'19) [23]	65.3	-	57.2
Use Temporal information			
Pavllo *et al*. (CVPR'19) [45]	64.7	-	-
Single-Image Method			
Li *et al*. (ICCV'19) [27]	88.8	-	66.5
Ours	**62.9**	**50.5**	**47.5**

Table 2: Comparison with SOTA weakly-supervised methods. Average MPJPE over all 15 actions for H36M under two protocols (P1 and P2) is reported. P1* refers to protocol 1 evaluated with ground truth 2d key-points. Best performance is marked with bold font. Error for each action can be found in our supplementary material.

Comparison with fully-supervised methods Here we compare with *fully-supervised methods* that uses the whole training split of H36M. We use S15678 as our initial population and Table 3 shows the performance comparison. Under this setting, our model also achieves competitive performance compared with other SOTA methods, indicating that our approach is not limited to scarce data scenario.

Method	**Performance**		
Authors	P1	P1*	P2
Martinez *et al*. (ICCV'17) [34]	62.9	45.5	47.7
Yang *et al*. (CVPR'18) [66]	58.6	-	**37.7**
Zhao *et al*. (CVPR'19) [68]	57.6	43.8	-
Sharma *et al*. (ICCV'19) [54]	58.0	-	40.9
Moon *et al*. (ICCV'19) [38]	54.4	35.2	-
Ours	**50.9**	**34.5**	38.0

Table 3: Comparison with SOTA methods under fully-supervised setting. Same P1, P1* and P2 as in Table 2.

5.3. Cross-dataset Generalization

To validate the generalization ability of our 2D-to-3D network in unknown environment, Table 4 compares with other methods on 3DHP. In this experiment we evolve from S15678 in H36M to obtain an augmented dataset consisting of 8 million 2D-3D pairs. Without utilizing any training data of 3DHP, we achieve SOTA performance in this

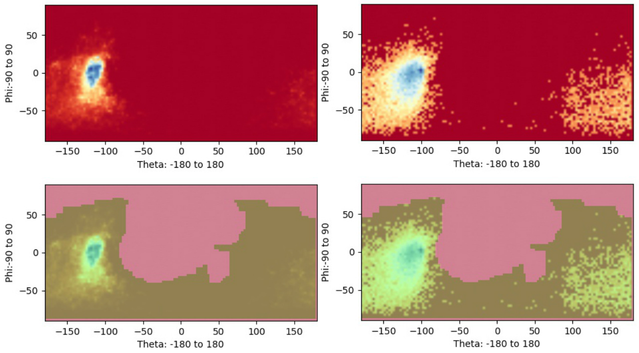


Figure 6: Dataset distribution for the bone vector connecting right shoulder to right elbow. Top: distribution before (left) and after (right) dataset augmentation. Bottom: distribution overlaid with valid regions (brown) taken from [2] .

benchmark. We obtain clear improvements comparing with [25], which also uses S15678 as the training data but fix it without data augmentation. The results indicate that our data augmentation approach improves model generalization effectively despite we start with the same biased training dataset. As shown in Fig. 6, the distribution of the augmented dataset indicates less dataset bias. Qualitative results on 3DHP and LSP are shown in Fig. 7. Note that these unconstrained poses are not well-represented in the original training dataset yet our model still gives good inference results. Qualitative comparison with [25] on some difficult poses in U3DPW is shown in Fig. 8 and our model shows better accuracy for these rare human poses.

Method	CE	PCK	AUC	MPJPE
Mehta et al. [35]		76.5	40.8	117.6
VNect [37]		76.6	40.4	124.7
LCR-Net [52]		59.6	27.6	158.4
Zhou et al. [69]		69.2	32.5	137.1
Multi Person [36]		75.2	37.8	122.2
OriNet [31]		81.8	45.2	**89.4**
Li et al. [25]	✓	67.9	-	-
Kanazawa [22]	✓	77.1	40.7	113.2
Yang et al. [66]	✓	69.0	32.0	-
Ours	✓	**81.2**	**46.1**	**99.7**

Table 4: Testing results for the MPI-INF-3DHP dataset. A higher value is better for PCK and AUC while a lower value is better for MPJPE. MPJPE is evaluated without rigid transformation. CE denotes cross-dataset evaluation and the training data in MPI-INF-3DHP is not used.

5.4. Ablation Study

Our ablation study is conducted on H36M and summarized in Table 5. The baseline (B) uses T=1. Note that adding cascade (B+C) and dataset evolution (B+C+E) consistently out-perform the baseline. Discussion on the evolution operators is included in our supplementary material.

Effect of cascade length T Here we train our model on various subsets of H36M and plot MPJPE over cascade length as shown in Fig. 9. Here R is fixed as 2. Note that the training error increases as the training set becomes more complex and the testing errors decreases accordingly. The gap between these two errors indicate insufficient training data. Note that with increasing number of deep learners, the training error is effectively reduced but the model does not overfit. This property is brought by the ensemble effect of multiple deep learners.

Effect of block number R Here we fix T=1, d=512 and vary R. S15678 in H36M and its evolved version are used. The datasets before (BE) and after evolution (AE) are randomly split into training and testing subsets for clarity. The training and testing MPJPEs are shown in Fig. 10. Note that the training error is larger after evolution with the same R=7. This means our approach brings novel information to the dataset, which can afford a deeper architecture with larger R (e.g. R=9).

Method	Training Data	P1	P1*
Problem Setting A: Weakly-supervised Learning			
B	S1	71.5	66.2
B+C	S1	$70.1_{\downarrow 2.0\%}$	$64.5_{\downarrow 2.6\%}$
B+C+E	Evolve(S1)	$62.9_{\downarrow 12.0\%}$	$50.5_{\downarrow 21.7\%}$
Problem Setting B: Fully-supervised Learning			
B	S15678	54.3	44.5
B+C	S15678	$52.1_{\downarrow 4.0\%}$	$42.9_{\downarrow 3.6\%}$
B+C+E	Evolve(S15678)	$50.9_{\downarrow 6.2\%}$	$34.5_{\downarrow 22.4\%}$

Table 5: Ablation study on H36M. B: baseline. C: add cascade. E: add data evolution. Evolve() represents the data augmentation operation. Same P1 and P1* as in Table 2. Error reduction compared with the baseline follows the ↓ signs.

6. Conclusion

This paper presents a novel evolution framework to enrich the data distribution of an initial biased training set, leading to better intra-dataset and cross-dataset generalization of 2D-to-3D network. A novel monocular human pose estimation model is trained achieving state-of-the-art performance for single-frame 3D human pose estimation. There are a lot of fruitful directions remaining to be explored. First, extension to temporal domain, multi-view setting and multi-person scenarios are just three examples. Second, instead of using fixed evolution operators, we will investigate how the operators can also evolve during the data generation process.

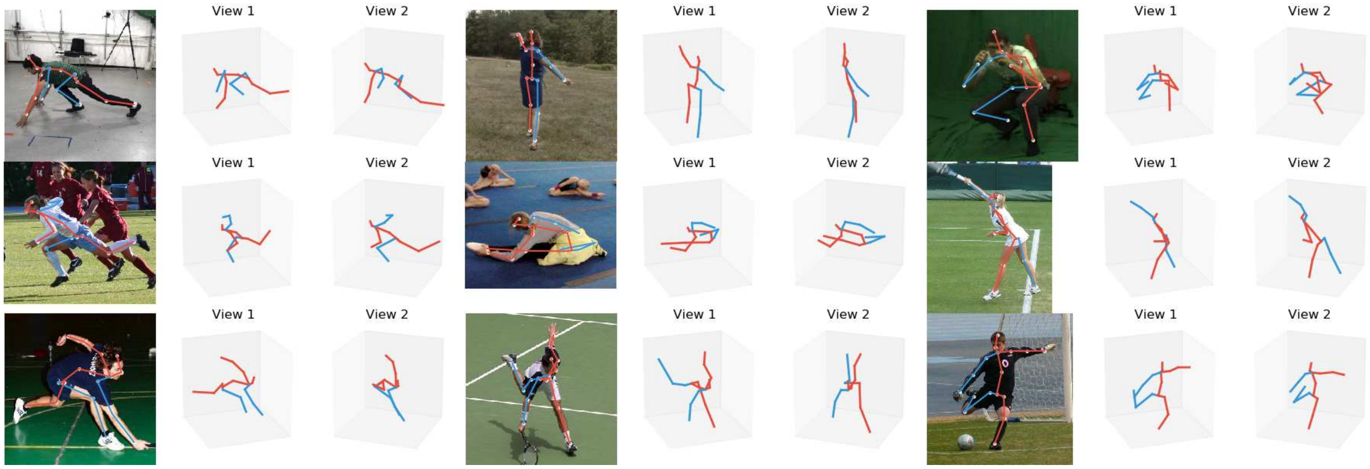


Figure 7: Cross-dataset inferences of $\mathcal{G}(\mathbf{c})$ on MPI-INF-3DHP (first row) and LSP (next two rows).

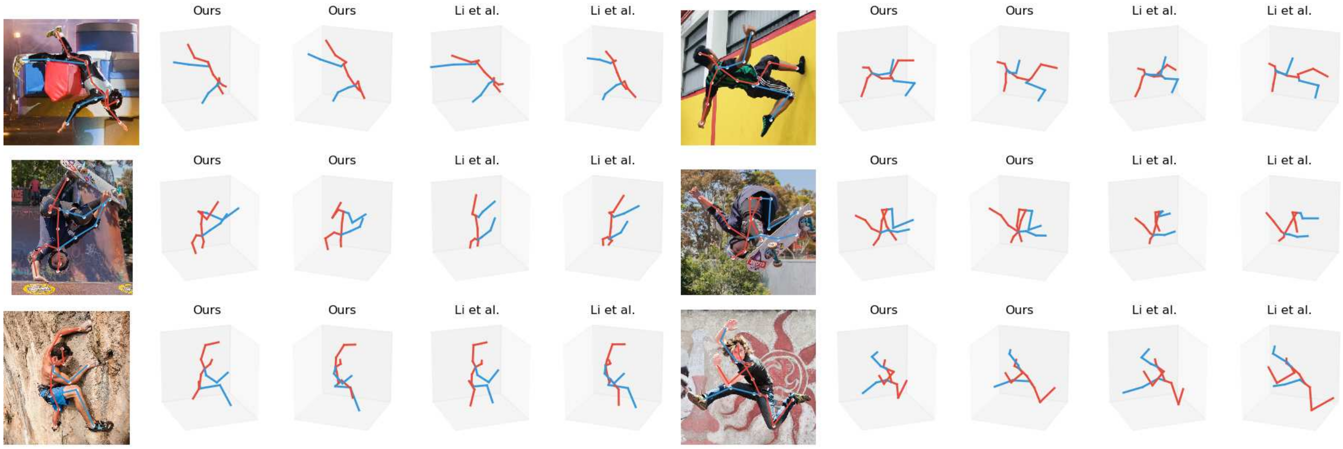


Figure 8: Cross-dataset inference results on U3DPW comparing with [25]. Video is included in our supplementary material.

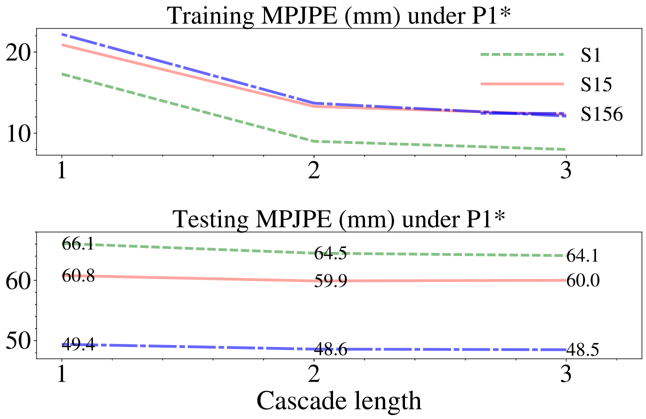


Figure 9: Training and testing errors with varying number of cascade length and training data. The cascade effectively reduces training error and is robust to over-fitting.

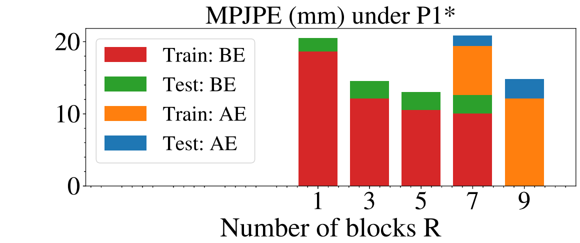


Figure 10: MPJPE (P1*) before (BE) and after evolution (AE) with varying number of blocks R. Evolved training data can afford a deeper network. Best viewed in color.

Acknowledgments We gratefully acknowledge the support of NVIDIA Corporation with the donation of one Titan Xp GPU used for this research. This research is also supported in part by Tencent and the Research Grant Council of the Hong Kong SAR under grant no. 1620818.

References

[1] Ankur Agarwal and Bill Triggs. Recovering 3d human pose from monocular images. *IEEE transactions on pattern analysis and machine intelligence*, 28(1):44–58, 2005.

[2] Ijaz Akhter and Michael J Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1446–1455, 2015.

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele.

[4] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM transactions on graphics (TOG)*, volume 24, pages 408–416. ACM, 2005.

[5] Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 3d pictorial structures for multiple human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1669–1676, 2014.

[6] Liefeng Bo and Cristian Sminchisescu. Structured output-associative regression. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2403–2410. IEEE, 2009.

[7] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision*, pages 561–578. Springer, 2016.

[8] Magnus Burenius, Josephine Sullivan, and Stefan Carlsson. 3d pictorial structures for multiple view articulated pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3618–3625, 2013.

[9] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014.

[10] Wenzheng Chen, Huan Wang, Yangyan Li, Hao Su, Zhenhua Wang, Changhe Tu, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Synthesizing training images for boosting human 3d pose estimation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 479–488. IEEE, 2016.

[11] Xipeng Chen, Kwan-Yee Lin, Wentao Liu, Chen Qian, and Liang Lin. Weakly-supervised discovery of geometry-aware representation for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10895–10904, 2019.

[12] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7103–7112, 2018.

[13] Yu Cheng, Bo Yang, Bo Wang, Wending Yan, and Robby T. Tan. Occlusion-aware networks for 3d human pose estimation in video. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[14] Hai Ci, Chunyu Wang, Xiaoxuan Ma, and Yizhou Wang. Optimizing network structure for 3d human pose estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[15] João Correia, Tiago Martins, and Penousal Machado. Evolutionary data augmentation in deep face detection. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pages 163–164, 2019.

[16] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Gerard Pons-Moll, and Christian Theobalt. In the wild human pose estimation using explicit 2d features and intermediate 3d representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10905–10914, 2019.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[18] John Henry Holland et al. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press, 1992.

[19] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[20] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013.

[21] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference*, 2010. doi:10.5244/C.24.12.

[22] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018.

[23] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Self-supervised learning of 3d human pose using multi-view geometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1077–1086, 2019.

[24] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[25] Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9887–9895, 2019.

[26] Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9572–9581, 2019.

[27] Zhi Li, Xuan Wang, Fei Wang, and Peilin Jiang. On boosting single-frame 3d human pose estimation via monocular videos. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[28] Mude Lin, Liang Lin, Xiaodan Liang, Keze Wang, and Hui Cheng. Recurrent 3d pose sequence machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 810–819, 2017.

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

[30] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):248, 2015.

[31] Chenxu Luo, Xiao Chu, and Alan Yuille. Orinet: A fully convolutional network for 3d human pose estimation. *arXiv preprint arXiv:1811.04989*, 2018.

[32] Diogo C Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5137–5146, 2018.

[33] Elisabeta Marinoiu, Dragos Papava, and Cristian Sminchisescu. Pictorial human spaces: How well do humans perceive a 3d articulated pose? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1289–1296, 2013.

[34] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2640–2649, 2017.

[35] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 International Conference on 3D Vision (3DV)*, pages 506–516. IEEE, 2017.

[36] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*, pages 120–130. IEEE, 2018.

[37] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4):44, 2017.

[38] Gyeongsik Moon, Juyong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *The IEEE Conference on International Conference on Computer Vision (ICCV)*, 2019.

[39] Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2823–2832, 2017.

[40] Aiden Nibali, Zhen He, Stuart Morgan, and Luke Prendergast. Numerical coordinate regression with convolutional neural networks. *arXiv preprint arXiv:1801.07372*, 2018.

[41] Bruce Xiaohan Nie, Ping Wei, and Song-Chun Zhu. Monocular 3d human pose estimation by predicting depth on joints. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3467–3475. IEEE, 2017.

[42] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019.

[43] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7025–7034, 2017.

[44] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Harvesting multiple views for marker-less 3d human pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6988–6997, 2017.

[45] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7753–7762, 2019.

[46] Xi Peng, Zhiqiang Tang, Fei Yang, Rogerio S Feris, and Dimitris Metaxas. Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2226–2234, 2018.

[47] Mir Rayat Imtiaz Hossain and James J Little. Exploiting temporal information for 3d human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 68–84, 2018.

[48] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1692, 2014.

[49] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3d human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 750–767, 2018.

[50] Helge Rhodin, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua. Learning monocular 3d human pose estimation from multi-view images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8437–8446, 2018.

[51] Grégory Rogez and Cordelia Schmid. Mocap-guided data augmentation for 3d pose estimation in the wild. In *Advances in neural information processing systems*, pages 3108–3116, 2016.

[52] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3433–3441, 2017.

[53] Rómer Rosales and Stan Sclaroff. Learning body pose via specialized maps. In *Advances in neural information processing systems*, pages 1263–1270, 2002.

[54] Saurabh Sharma, Pavan Teja Varigonda, Prashast Bindal, Abhishek Sharma, and Arjun Jain. Monocular 3d human pose estimation by generation and ordinal ranking. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[55] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.

[56] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1-2):4, 2010.

[57] William M Spears. Crossover or mutation? In *Foundations of genetic algorithms*, volume 2, pages 221–237. Elsevier, 1993.

[58] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[59] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019.

[60] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 529–545, 2018.

[61] Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. A deeper look at dataset bias. In *Domain adaptation in computer vision applications*, pages 37–55. Springer, 2017.

[62] A Torralba and AA Efros. Unbiased look at dataset bias. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1521–1528. IEEE Computer Society, 2011.

[63] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 109–117, 2017.

[64] Bastian Wandt and Bodo Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[65] Bastian Wandt and Bodo Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7782–7791, 2019.

[66] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5255–5264, 2018.

[67] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2248–2255, 2013.

[68] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3425–3435, 2019.

[69] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 398–407, 2017.

[70] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4966–4975, 2016.