A Synthesis Approach to Semi-Supervised 3D Pose Estimation from Single Images

Anonymous CVPR submission

Paper ID 1818

Abstract

We present a new semi-supervised learning approach for estimating 3D poses from single images. Our method consists of two learning stages: unsupervised and supervised learning. Notably, **no human-annotated 3D data** *is required even for training deep network under full supervision, making it easy to transfer from human poses to animal poses estimation.*

Our unsupervised reconstruction network is an autoencoder trained with skip connections for progressively learning complex poses from images. In the supervised stage, interleaved layers are learned to produce 2D joint landmarks using synthetic images. Then, we generate a massive synthetic training dataset containing more than 130 million ground-truth 3D poses with corresponding 2D landmarks. Thus our data acquisition is unbiased and totally markerless, in contrast to representative 3D pose datasets requiring custom capture on a limited number of human actor subjects. The 3D synthetic poses are valid since they are constrained by a hierarchical 3D skeletal model with anatomical joint constraints. Quantitative results show that our model achieves the state-of-the-art performance, and qualitative results show that our model is capable of estimating complex and extreme poses.

1. Introduction

This paper presents a new semi-supervised learning approach for *3D* pose estimation from single images, using unlabeled data for unsupervised training and *no* human-annotated 3D data for training deep models in full supervision. Pose estimation from an RGB image is an important task in computer vision, which can be applied to behavior analysis, action recognition [1] and so on. With the availability of massive 2D human pose datasets, such as MPII [2] and COCO [3], by utilizing deep convolution network with special designs, the state-of-the-art methods achieve high accuracy in 2D single/multiple human pose estimation, even when the input image exhibits significant occlusion.

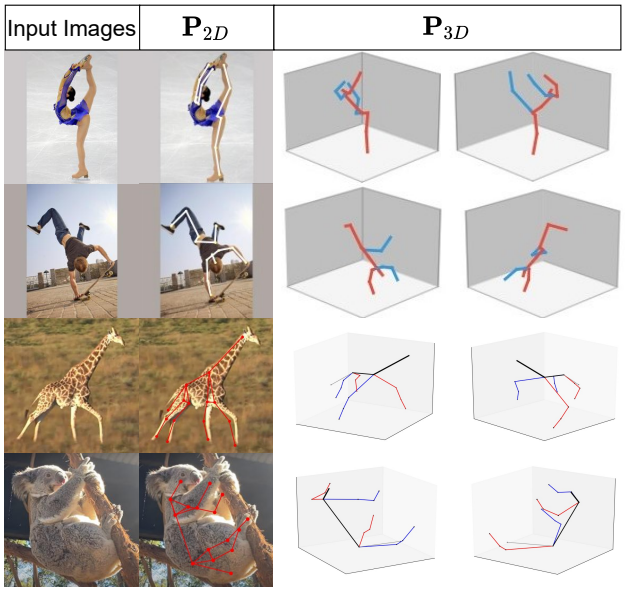

Figure 1. 3D pose estimation results from single images. For humans, our model can estimate extreme poses *without* ground-truth annotations. For animals, our model can adapt to species with different limb and poses with only a *few* 2D ground-truth annotations.

However, it is prohibitively laborious to manually annotate 3D pose from RGB images in large quantity, making learning 3D human pose estimation from single RGB images a very challenging problem in computer vision. Existing large-scale 3D pose datasets are limited in availability and often biased to a small number of human subjects performing limited classes of actions captured under laboratory settings [4], except for [5] whose model was trained using a large-scale synthetic dataset mapping perfect 2D landmarks (not real landmarks detected from images) to 3D human poses. Models trained with biased data are liable to fail in cross-dataset validation.

Inspired by [6], we propose a semi-supervised learning approach to infer 3D human poses from single RGB images which, similar to [5] requires no laborious ground-

truth annotations while achieving high cross-dataset generalization ability even applicable to estimate 3D extreme poses. In summary, we propose unsupervised learning using a reconstruction network followed by fully-supervised learning trained using synthetic datasets:

Our unsupervised reconstruction network consists an autoencoder augmented with interleaved layers in the decoder stage for transfer to 2D landmark detection. This is inspired by [6] which was designed for reconstructing mostly frontal faces and fails in learning complex structures of human poses. To upgrade this network, we incorporate progressive training using skip-connection layers to enable the reconstruction network to learn complex poses from human images. The reconstruction network can be trained by a large collection of *unlabeled* human images.

In the supervised stage, the interleaved layers can be trained by a small number of hand-annotated images or *synthetic* human images. The implicit knowledge learned by the reconstruction network enables the model to generalize from synthetic data to real-world data. Finally, with a deep model trained using massive, unbiased 3D synthetic poses generated on-the-fly at 500Hz without overfitting[1] in the full supervision stage, we can easily generalize the whole semi-supervised model to estimate human and animal poses, as well as others (e.g. hand) in the future. Figure 1 shows complex 3D pose estimation results for humans and animals.

2. Related Work

2.1. Human Pose Estimation

2D Pose Estimation. Deep learning on pose estimation was first proposed in [7] where convolutional neural network was used for regressing body joints. An efficient position refinement model was proposed in [8]. Convolutional pose machine [9] consists of an image feature computational module along with a prediction module. A self-correcting model was used in [10] that progressively refines the initial prediction by feeding back errors. The well-known landmark detector [11] performs repeated bottom-up, top-down processing with intermediate supervision. In [12] the authors presented an approach based on the ResNet appended with a number of deconvolutional layers. In [13] parallel branches were proposed to deal with images in high resolution. In [14, 15] various approaches were proposed to refine heatmap regression. In [14, 13, 15] state-of-the-art results on 2D key-point detection and single/multiple person pose estimation on COCO is reported [3].

3D Pose Estimation. Multi-task learning framework was used in [16] for joint point regression and detection to disentangle dependencies among different body parts while learning their correlations. In [17], numerous 2D projections were generated with depth estimated by traversing a large 3D pose library. 2D and 3D labels were used in [18] as input to a two-stage cascaded structure. Integral operation was used in [19] to relate and unify heat map representation and joint regression. Recent research focuses on multi-person pose estimation from multiple viewpoints [20], or weakly-supervised training to analyze images in the wild [21].

More recently, in [22] 3D poses and mesh were estimated from single images. Weakly supervised 3D human shape and poses from low-resolution single images were studied in [23]. Part-aware [24] and local parts [25] were employed to respectively search for optimal architecture and generalize to unseen poses. Sophisticated model has been proposed to estimate 3D humans pose jointly with camera, appearance and part segmentation [26]. All the above recent works use existing human-annotated data data such as [4] for training, and are thus vulnerable to biases and limitation to the human subjects and motion capture settings. In [5, 27, 28], a human parametric model was used to lift 2D poses to 3D. But these methods require accurate detection of 2D poses where the relevant model has to be trained by massive training data. In this work, almost human cost-free synthetic data are used to train the 2D landmark detector for humans and animals.

Human Pose Datasets. Common pose datasets such as MPII [2], LSP [29], PoseTrack [30] and FLIC [31] are 2D human pose datasets with annotated body joints. Each of them contains around only 10K images. Human 3.6M [4] contains 3.6 million 3D human poses with their corresponding images captured from 17 professional actors. Notable 3D human datasets include HumanEva [32], Monocular 3D [33], Unite the People [34], DensePose [35], SURREAL [36] and VGG [37]. In this paper, pose-guided person image generation [38, 39, 40, 41, 42] are deployed to synthesize human images in the unsupervised learning stage, without laborious human annotation for unseen action classes which are difficult to capture.

2.2. Animal Pose Estimation

Recent works have contributed to transferring human to proximate animal poses [43] and estimating insect poses [44]. It can be impractical or even deadly to capture real 3D animal poses, which is not an issue in our synthesis approach. Biologists had used the DeepLabCut toolbox [45] to mark feature points on animals in a video. Transfer learning from human to animal pose is applied. Their toolbox requires a large amount of human annotation, and most of their videos were taken under special laboratory setting and environment. A number of applications was proposed in their subsequent work [46]. Although they also used a rough animal skeleton for analysis, such prediction

[1] Trained for 3 days or 259,200 seconds at a rate of seeing 500 poses per second, our model had so far seen over 130 million valid 2D-to-3D poses.

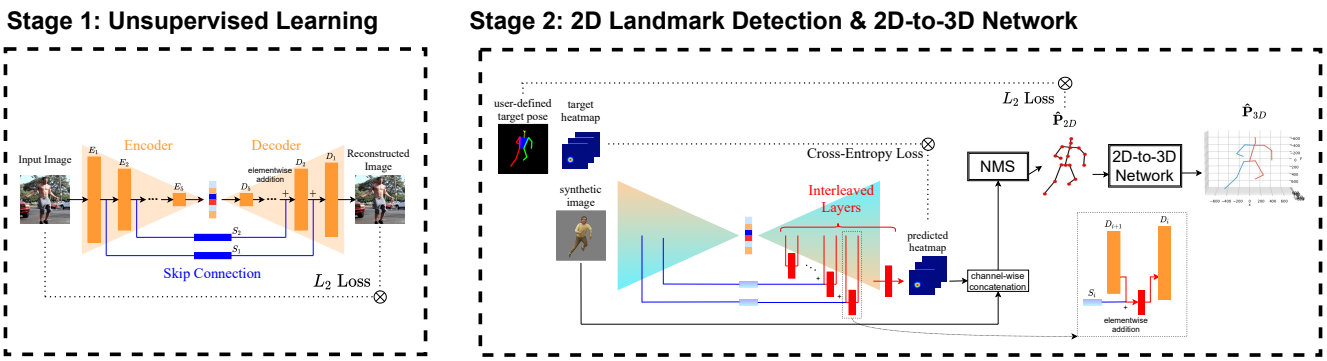

Figure 2. Overview of architecture and training. Stage 1: unsupervised training of autoencoder with our skip connections. The skip connections are progressively trained to improve human images reconstruction. Stage 2: supervised training of interleaved layers from [6] for 2D joint heatmaps, non-maximal suppression for 2D joints $\mathbf{P}_{2D}$, and 2D-to-3D network for 3D poses $\mathbf{P}_{3D}$. For the 2D-to-3D network, [5] is used for human pose estimation, while our own 2D-to-3D network is used for animal pose estimation.

was based on pictures taken from six different cameras and had very limited set of joints on the graph.

Animal Pose Dataset. Recent work has also contributed a large-scale dataset of animals annotated with facial landmarks [47]. Most common datasets such as ImageNet [48] and YouTube8M [49] contain labeled animals images/videos but they lack the necessary joint location information. Tigdog [50] provides annotated landmarks but its size is not comparable to the above-mentioned human pose datasets. To generate an animal pose dataset that is sufficiently complete (sufficient coverage in the articulated motion space) and compact (no invalid poses), applicable range of motion for each joint should be applied. Different animals have different valid ranges, and we refer readers to pertinent anatomical works on quadrupedal mammals, such as the range of head-neck movement [51], arm glenoid line [52] and knee joint [53], which guide the setting on the parameter range in constructing our synthetic animal pose dataset.

3. Method

Figure 2 shows the overall network architecture. Though apparently simple, this produces state-of-the-art results using the massive, unbiased and valid 3D human/animal poses *synthesized on-the-fly* for training deep networks.

Our model is trained in two stages, namely, unsupervised and supervised learning. In unsupervised learning, massive unlabeled human images are used to train the encoder-decoder with progressive training of the skip connection layers. Synthetic pose-guided images or a manageable number of hand-annotated images are used to train the interleaved layers to output 2D pose heatmap, which is then fed to a non-maximum suppression module to output 2D landmarks. These landmarks are in turn fed as input to a 2D-to-3D network which is trained in full supervision using massive synthetic 3D poses.

3.1. Unsupervised Image Reconstruction

Encoder-Decoder Network. A large-scale, unlabeled human image dataset is used to train our human image reconstruction network without supervision. The implicit poses knowledge inherent in the unlabeled human images learned by the reconstruction network provides the generalization capability of the 2D landmark detector. Our encoder-decoder is directly derived from ResNet (see supplementary material for detailed architecture). All encoder and decoder layers are optimized using L_2 loss between the input and reconstructed images. Roughly 90K center-cropped human pose images from PoseTrack [54] are used in training. Standard encoder-decoder training is applied:

$$\min_{\mathbf{E},\mathbf{D}} ||I - \mathbf{D}(\mathbf{E}(I))||_2^2 \tag{1}$$

where I is the input image, and $\mathbf{E}$, $\mathbf{D}$ are respectively the encoder and decoder.

Since autoencoders have multiple convolutional and deconvolutional layers, they are liable to information loss in image reconstruction. To improve the performance, skip connections can be added from the encoder to the decoder across the bottleneck.

Skip Connection. In [6] human facial images were successfully reconstructed in high quality. However the network falls short of reconstructing general human images because they are far more complex than facial images. Facial or portrait images exhibit little occlusion and has simpler geometry making training and testing easier. For more complex human images, the reconstruction loss would converge to an unsatisfactory point due to the inherent drawback of L_2 loss: same weights to pixels regardless they belong to high or low frequency features. Consequently, blurry reconstructed human images are resulted.

As shown in Figure 2 we insert skip connections, which are 3×3 same-padding convolution layers with equal input and output shape. All skip connection layers share

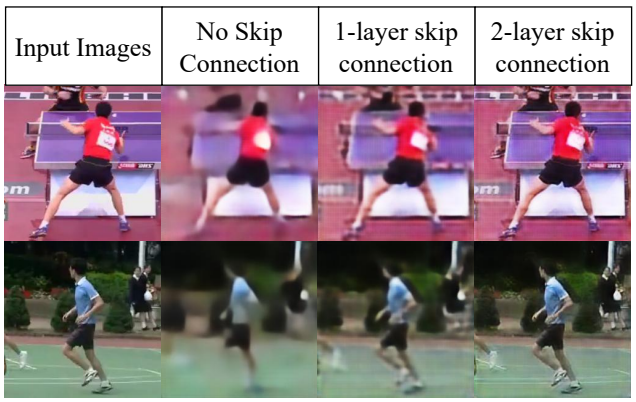

Figure 3. Comparison of the reconstructed images before and after applying skip connection.

the same structure. Our skip connections are trained progressively proceeding from inner to outer layers. Figure 3 shows the progressive training results in stages. Freezing all the encoder-decoder layers trained aforementioned, we start from training the inner-most skip connection, i.e., the skip connection between the last encoder layer and the first decoder layer. Upon convergence, we freeze the well-trained first skip connection and continue to train the second skip connection, so on. Such progressive training prevents the skip connection from directly passing the image information to the reconstructed images, which will lead to undesired optimum that the skip connection outputs identity and the encoder-decoder network outputs zero. With progressive training L_2 loss can now be used for optimizing the skip connections.

In our experiment, we found that the inner skip connection layers converge very slowly, while outer skip connection layers suffice to reduce the reconstruction loss. Therefore, we only add skip connection between the two outer-most encoder-decoder layers which are progressively trained.

3.2. Supervised 2D Landmarks Transfer

Interleaved Layers. With faithful reconstruction of complex human images, the model can be transferred to 2D pose/landmark estimation similar to style transfer [6] with fewer parameters to estimate. The output in this stage is a landmark heatmap.

During this training stage, the encoder-decoder network with skip connections are frozen to constrain the training of the interleaved layers (Figure 2) and prevent them from overfitting. The interleaved layers are 3×3 same-padding convolution layers with equal input and output size. They are inserted across the decoder layers as shown. For example, the first decoder layer and its corresponding skip connection layer will output two matrices with the shape $512 \times 4 \times 4$. The two matrices are fed into two 3×3 same-padding convolution and two $512 \times 4 \times 4$ matrices are obtained. The two matrices are element-wise added and fed into the next decoder layer. This is similarly done for other interleaved layers.

We optimize the interleaved layers using the weighted cross entropy loss between target heatmaps $\mathbf{Y}$ and predicted heatmaps $\hat{\mathbf{Y}}$, freezing the encoder-decoder layers and the skip connection:

$$\mathcal{L}(\mathbf{Y}, \hat{\mathbf{Y}}) = \lambda \mathbf{Y} \cdot \log(\hat{\mathbf{Y}}) + (\mathbf{1} - \mathbf{Y}) \cdot \log(\mathbf{1} - \hat{\mathbf{Y}}) \quad (2)$$

Since most of the entries in the target heatmaps are nearly zeroes, using naive cross-entropy loss (i.e., $\lambda = 1$) will predict a zero heatmap without inducing much loss, which leads to slow convergence of the training. Therefore, the weight λ is added to penalize the error made when the target heat values are close to 1. In our model, the target heatmap is created using Gaussian distribution with $\sigma = 10$. The λ is chosen to be $256^2/\sigma^2\pi \approx 208.7$ to yield fast convergence.

Non-Maximal Suppression. While working for faces [6], straightforward argmax does not work for transforming heatmaps to complex 2D pose landmarks. We observe that the estimated heatmap can propose several likely candidates for a keypoint, but the correct one does not always has the highest heat value. Thus, we train a non-maximal suppression module to produce the most likely landmark, given the input image and estimated heatmap.

The non-maximal suppression module is a standard ResNet, receiving channel-wise concatenation of the input image and the estimated heatmap as input, and outputting normalized coordinates. This network has the same architecture as the encoder, except a sigmoid module is appended for outputting the normalized coordinates. The input shape is $17 \times 256 \times 256$, where $17 = 3$ (RGB) $+14$ (predicted joints). The output size is 28 (#joints $\times 2$). L_2 loss between the real coordinates and the predicted coordinates is used to optimize this heatmap-to-landmark network.

3.3. 2D Landmarks to 3D Poses

After estimating the 2D landmarks from non-maximal suppression, a 2D-to-3D network is used to lift the pose from 2D to 3D. Without losing generality, in the following we describe our technical contributions for estimating general 3D animal poses (human is a special case), as we can simply replace unlabeled human images by unlabeled animal images in the above to obtain 2D animal pose landmarks.

Here, we adopt a 3D hierarchical skeletal model similar to [5] which is commonly used in 3D character animation. Such 3D skeletal model can effectively constrain the estimated joint length and angles (parameters) within the valid pose space. This parametric approach also enables easy generation of our massive data set which consists of more than 25 millions ground-truth animal 3D poses with corresponding 2D landmarks. Sufficient training data is thus

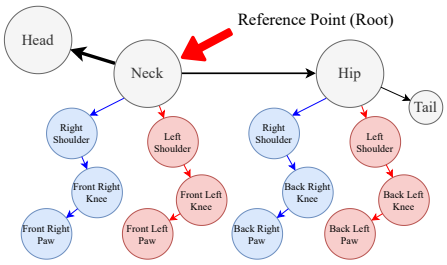

Figure 4. The hierarchical model contains 16 nodes roughly in 5 parts: backbone, front left limb, front right limb, back left limb and back right limb. The absolute location of each joint is recursively retrieved from the relative position with its parent node in local coordinates.

available for a deep network to learn how to resolve complex pose ambiguities and self-occlusions due to 2D projections. Thus it is easy to extend our network to estimate other 3D poses (e.g., hand gestures) using unlabeled hand images and synthetic poses generated using a similar simple skeletal model:

3.3.1 Parametric Skeletal Model

Our animal skeletal model is a tree encompassing 16 nodes with 15 edges. Figure 4 shows the skeletal model, which is rooted at the neck. $\mathbf{P}_{3D}$ denotes the 3D pose in global coordinates, which is a 3×16 matrix where each column represents the cartesian coordinates of each node:

$$\begin{aligned} \mathbf{P}_{3D} &= [P_1, P_2, \cdots, P_{16}] \\ P_i &= [x_i, y_i, z_i]^T \end{aligned} \tag{3}$$

We use $\mathbf{O}$ for edges, which consists of 15 pairs in spherical coordinates:

$$\begin{aligned} \mathbf{O} &= (O_1, O_2, \cdots, O_{15}) \\ O_i &= (r_i, \theta_i, \phi_i) \end{aligned} \tag{4}$$

where r, θ, ϕ are the (spherical) coordinates with respect to its parent node. With the hierarchical model, given the 3D coordinates of the parent node (x_p, y_p, z_p), the 3D coordinates of its child node (x_i, y_i, z_i) can be determined as:

$$\begin{aligned} x_i &= x_p + r_i \sin\theta_i \cos\phi_i \\ y_i &= y_p + r_i \sin\theta_i \sin\phi_i \\ z_i &= z_p + r_i \cos\theta_i \end{aligned} \tag{5}$$

which can be applied recursively in the tree structure so that the local coordinates $\mathbf{O}$ can be converted to the global coordinates $\mathbf{P}_{3D}$. Conversely, the global coordinates $\mathbf{P}_{3D}$ can also be converted to the local coordinates $\mathbf{O}$ by transforming the cartesian coordinates to the spherical coordinates. Hence, the transformation f between $\mathbf{P}_{3D}$ and $\mathbf{O}$ are:

$$\mathbf{P}_{3D} = f(\mathbf{O}) \quad \mathbf{O} = f^{-1}(\mathbf{P}_{3D}) \tag{6}$$

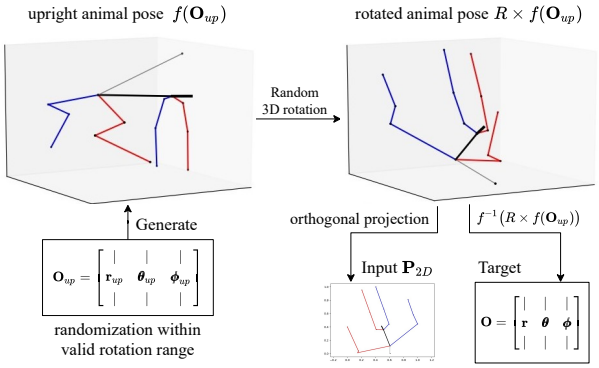

Figure 5. Workflow of synthesizing a valid 3D animal pose and the construction of the training inputs and the training targets for the 2D-to-3D network; the f in the figure is the transformation in Equation 6.

3.3.2 Synthetic Pose Dataset

As shown in Figure 5, we empirically define the valid rotation range of each edge when the animal is standing upright and facing negative-y axis. Based on the empirically valid rotation range, random values are assigned to (θ_i, ϕ_i). For r_i, all the edge lengths are randomly chosen following the uniform distribution between 0 and 1, and then are normalized so that the longest edge has unit length. After this random synthesis of a valid upright animal, random 3D rotation is applied so that the synthetic data can cover all possible animal poses (e.g. climbing trees, lying on the ground). Finally, the training inputs $\mathbf{P}_{2d}$ of the 2D-to-3D network are obtained by the projection of the synthetic 3D poses onto the xz-plane. The training targets $\mathbf{O}$ are the local coordinates of the 3D poses:

$$\begin{aligned} \mathbf{P}_{2D} &= P_{xz} \cdot R \cdot f(\mathbf{O}_{\text{up}}) \\ \mathbf{O} &= f^{-1}\big(R \cdot f(\mathbf{O}_{\text{up}})\big) \end{aligned} \tag{7}$$

where $\mathbf{O}_{\text{up}}$ is the local coordinate representation of the synthetic upright animal pose, f is the transformation defined in Equation 6, R is the randomized 3D rotation matrix, P_{xz} is the orthogonal projection matrix onto the xz-plane, and $\cdot$ denotes matrix multiplication.

An ideal synthetic dataset should be both complete (include all possible valid poses) and compact (exclude all invalid poses). We argue that both completeness and compactness are achieved to a high degree by: 1) the uniform randomization scheme of r_i which exhausts all species with different limbs length, 2) the randomized local coordinates and the random 3D rotation which exhaust all possible animal poses, together with 3) the constraints imposed by the skeletal model which exclude all invalid animal poses.

3.3.3 2D-to-3D Network

The input to the network is a 32 (2×16 nodes) dimensional vector which is the flattened 2D pose $\mathbf{P}_{2D}$; the net-

work then outputs a 45 (3×15 edges) dimensional vector which is the flattened 3D local coordinates $\hat{\mathbf{O}} = (\hat{\mathbf{r}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}})$. Figure 6 shows the main component of the estimation network which consists of three cascaded residual blocks, where each block has two fully connected layers each of size 1024 followed by batch normalization [55] and RELU activation [56]. In addition, one fully connected layer is applied before the residual blocks to increase the input dimensionality to 1024, and the other one is applied before the final prediction to reduce the size to 45. For the 15 output neurons corresponding to $\{\hat{r}_i\}_{i=1}^{15}$ in the local coordinates, a sigmoid module is appended to enforce outputting normalized edge lengths.

With our parametric representation, we define the following losses to guarantee only valid poses are generated to appropriately measure the difference among poses. Let (r, θ, ϕ) denote the target and $(\hat{r}, \hat{\theta}, \hat{\phi})$ the prediction, then:

Symmetry Loss. Due to the symmetrical structure of animals, the left and right limbs should have the same length, which can be translated into:

$$\mathcal{L}_{sym} = \frac{1}{|\ell|} \sum_{(i,i') \in \ell} (\hat{r}_i - \hat{r}_{i'})^2 \tag{8}$$

where ℓ is the subset of symmetric joint pairs (e.g., left and right limbs) and (i, i') are the corresponding left and right body parts. $|\ell|$ denotes the number of symmetric pairs.

Edges Loss. Negative cosine similarity between the target edges and the predicted edges is used to optimize the estimation of θ_i and ϕ_i:

$$\mathcal{L}_{ed} = -\frac{1}{15} \sum_{i=1}^{15} \begin{bmatrix} \sin\theta_i \cos\phi_i \\ \sin\theta_i \sin\phi_i \\ \cos\theta_i \end{bmatrix} \bullet \begin{bmatrix} \sin\hat{\theta}_i \cos\hat{\phi}_i \\ \sin\hat{\theta}_i \sin\hat{\phi}_i \\ \cos\hat{\theta}_i \end{bmatrix} \tag{9}$$

L_2 loss between the target edge lengths and the estimated edge lengths is used to optimize the estimation of r_i:

$$\mathcal{L}_{el} = \frac{1}{15} \sum_{i=1}^{15} (r_i - \hat{r}_i)^2 \tag{10}$$

Final loss function. Combining the above, the final loss function is:

$$\mathcal{L} = \mathcal{L}_{sym} + \mathcal{L}_{el} + \mathcal{L}_{ed} \tag{11}$$

Since the scales of the three losses are the same, addition without weights is sufficient to yield even convergence of the three losses.

4. Implementation

Training. All images used in this work are 256×256 with the output heatmap having the same resolution. The Adam optimizer [57] is used in all optimizations. The unsupervised training codes are run on an NVIDIA Titan GPU with learning rate $1e^{-4}$. We train the encoder-decoder for 48 hours to reduce the L_2 loss to 0.005. The total number of unlabeled human images is 90K. We train the skip connection layers progressively from S_2 to S_1 (Figure 2). For each layer of skip connection, we train for another 5 epochs after the L_2 loss has reduced to 0.005.

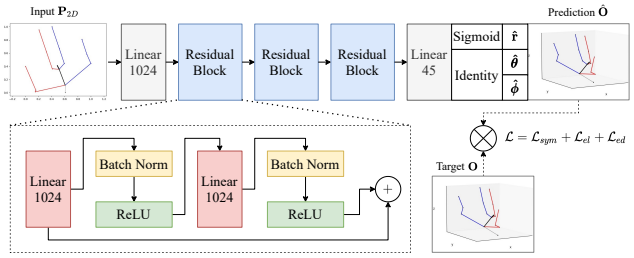

Figure 6. Our 2D-to-3D network. $\mathbf{P}_{2D}$ and $\hat{\mathbf{O}} = (\hat{\mathbf{r}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}})$ are respectively the concatenated input and output of the network as described in the text and Figure 5. The $\mathcal{L}$ is the loss defined in the Equation 11.

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	PCK ↑
Chu et al. [58]	98.1	93.7	89.3	86.9	93.4	94.0	92.5	92.6
Ours	89.3	86.4	74.3	59.9	96.1	86.0	74.4	81.3

Table 1. 2D landmark estimation accuracy on the LSP dataset [29]. Mean PCK is shown in the last column.

For the markerless 2D landmark detection, we train until the weighted cross-entropy loss between the targets and predictions reduces to $1e^{-4}$. Then we freeze the reconstruction network and the interleaved layers, and train the non-maximal suppression module until the landmark loss reduces to $1e^{-3}$. The learning rate in this step is $1e^{-4}$.

For the training of the 2D-to-3D network, a batch of 500 synthetic animal poses is used for one step of the Adam optimization, with another 500 synthetic poses used for validation. All the synthetic poses are used only once and discarded immediately. The training can in theory last forever without overfitting since we have infinite synthetic animal poses. We train for 50K batches, which means our model has seen 25 million animal poses. After training for 50K batches with the learning rate $1e^{-6}$, the validation loss will be respectively reduced to $\mathcal{L}_{sym} \approx 0.001$, $\mathcal{L}_{ed} \approx -0.98$, $\mathcal{L}_{el} \approx 0.01$.

Automatic Supervision vs. Fine-Tuning. With our 16-joint parametric skeletal model, although the empirically valid rotation range defined at the beginning may not cover unusual or extreme poses during testing, it can be easily solved by synthesizing a range of roughly similar 3D poses almost for free by re-defining the valid rotation range.

This is in stark contrast to traditional 2D-to-3D pose estimation trained using existing large-scale (but limited) 3D datasets which requires fine-tuning to learn unusual poses not seen during training. Such fine-tuning can be very tedious, because it requires labeling of *accurate 3D poses* corresponding to 2D landmarks similar to the new and unseen pose data in testing, which can often only be achieved by laborious human annotation.

Method Authors	WS	MPJPE ↓ P1	P2
Martinez et al. (ICCV'17) [59]	✗	62.9	47.7
Yang et al. (CVPR'18) [60]	✗	58.6	37.7
Zhao et al. (CVPR'19) [61]	✗	57.6	-
Sharma et al. (ICCV'19) [62]	✗	58.0	40.9
Moon et al. (ICCV'19) [63]	✗	54.4	-
Li et al. (CVPR'20) [5]	✗	49.7	38.0
Use Multi-view			
Rhodin et al. (CVPR'18) [64]	✓	-	64.6
Kocabas et al. (CVPR'19) [65]	✓	65.3	57.2
Use Temporal information			
Pavllo et al. (CVPR'19) [66]	✓	64.7	-
Single-image method			
Li et al. (ICCV'19) [67]	✓	88.8	66.5
Li et al. (CVPR'20) [5]	✓	62.9	47.5
Ours	✓	90.4	62.8

Table 2. Comparison with the SOTA methods. The performance is measured by the MPJPE over the 15 actions, under two protocols P1 and P2 (see Section 5.2). WS indicates whether the method is weakly-supervised. Li et al. [5] which uses perfect synthetic 2D landmarks is the upper bound of our method.

5. Experiments

5.1. 2D Landmark Estimation

Since our unsupervised 2D landmark detection is an intermediate step to achieve the final goal of 3D pose estimation, we only compare with the fully-supervised state-of-the-art method, focusing more on the evaluation of the 3D pose estimation in the following. Table 1 tabulates the accuracy of our model on the LSP dataset [29]. We use the Percentage of Correct Keypoints (PCK) measure at 0.2 with PC annotations as the evaluation metrics.

5.2. 3D Human Pose Estimation

Metrics We use *Mean Per Joint Position Error* (MPJPE) measured in millimeters, and *Percentage of Correct Keypoints* (PCK) which measures the correctness of the prediction under a specific threshold, and *Area Under the Curve* (AUC) which is the area under the correctness-threshold curve as the evaluation metrics. We directly adopted Li et al. [5] (marked with grey in Tables 2 and 3) as our 2D-to-3D network which uses ground truth 2D landmarks as input, making it the upper-bound of our method.

Human 3.6M (H3.6M) [4] is one of the largest 3D human pose datasets, which captures 15 regular daily activities from four camera views performed by 11 actors. The ground truth 3D poses are obtained using motion capture system, and the 2D landmarks are obtained by projection of the 3D poses based on the camera parameters. MPJPE is used as the performance metric for this dataset. Two protocols are applied for the evaluation. Protocol 1 (P1) directly computes the MPJPE. Protocol 2 (P2) computes the MPJPE after an alignment of the predicted pose.

Method	CE	PCK ↑	AUC ↑	MPJPE ↓
Mehta et al. [33]	✗	76.5	40.8	117.6
VNect [68]	✗	76.6	40.4	124.7
LCR-Net [69]	✗	59.6	27.6	158.4
Zhou et al. [18]	✗	69.2	32.5	137.1
Multi Person [70]	✗	75.2	37.8	122.2
OriNet [71]	✗	**81.8**	**45.2**	**89.4**
Li et al. [67]	✓	67.9	-	-
Kanazawa [72]	✓	**77.1**	**40.7**	113.2
Yang et al. [60]	✓	69.0	32.0	-
Li et al. [5]	✓	81.2	46.1	99.7
Ours	✓	75.2	39.8	**106.9**

Table 3. Results for the MPI-INF-3DHP dataset. Higher values of PCK and AUC are better and lower values of MPJPE are better. CE indicates the cross-dataset evaluation, i.e. no training data in MPI-INF-3DHP is used. [5] is the upper bound of our method.

H3.6M provides 7 subsets of data, which are denoted as S1, S5, S6, S7, S8, S9, S11. We compare our method with the weakly supervised methods and the fully supervised methods. We trained the fully-supervised models on S15678 and trained the weakly-supervised models on S1[2]. All the models were tested on S9 and S11. Table 2 shows the comparison experiments where weakly-supervised single-image methods should be directly compared. In particular, we use a simple method to synthesize massive and unbiased 3D poses for training and are comparable to [67] which trains using H3.6M and generates multiple 3D pose hypotheses using a more complicated multi-modal mixture density model.

MPI-INF-3DHP [33] is a benchmark to evaluate cross-dataset generalization capability. Since our model is not trained on the training set provided by MPI-INF-3DHP, the evaluation of our model is cross-dataset. MPJPE, PCK, AUC are used as the performance metrics for this dataset. Table 3 shows the comparison experiment: we are ranked second in cross-dataset evaluation after [72] where a generative 3D human body model is used. In [60], a generative adversarial network is proposed with sophisticated multi-source discriminator to distinguish invalid poses where 2D human annotation is still required.

Qualitative Results on LSP [29]. Figure 7 shows qualitative results on extreme poses estimation on LSP. Our model can generalize to extreme human poses by simply synthesizing the poses without any ground truth 3D annotation.

5.3. Animal Pose Estimation

We evaluate our model quantitatively on existing animal datasets [50] by feeding less than three ground-truth

[2]The experiment setting is from [5] where the comparisons can be directly made by rescaling the loss using a factor of 1.97 under the same setting. Refer to the supplementary material for more details of synthesis.

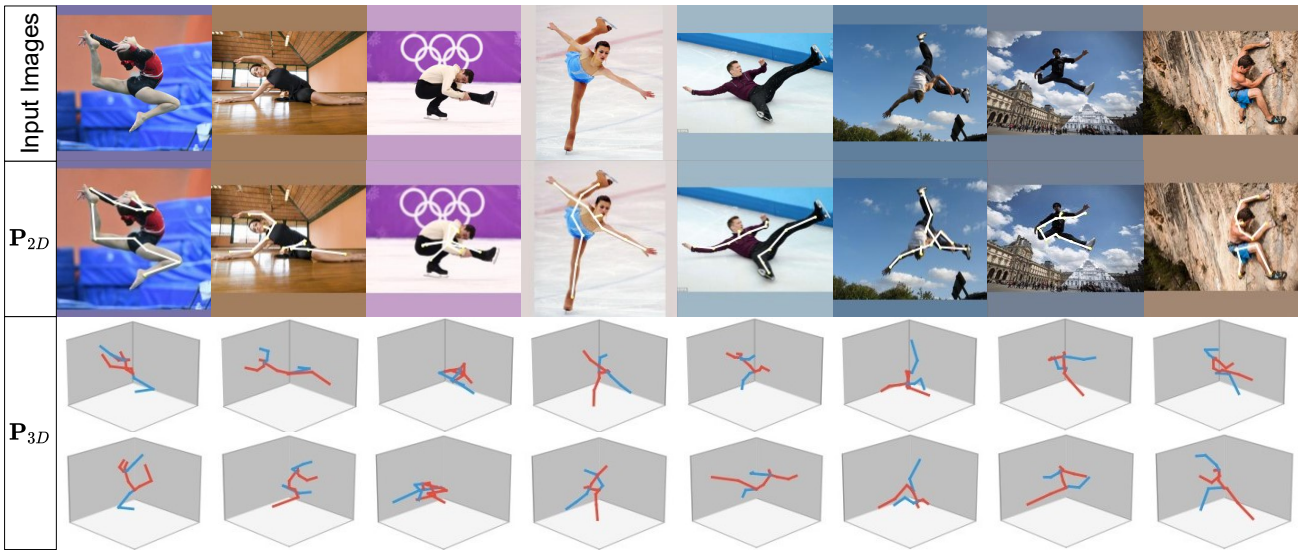

Figure 7. Human qualitative results on LSP. More results can be found in the supplementary material.

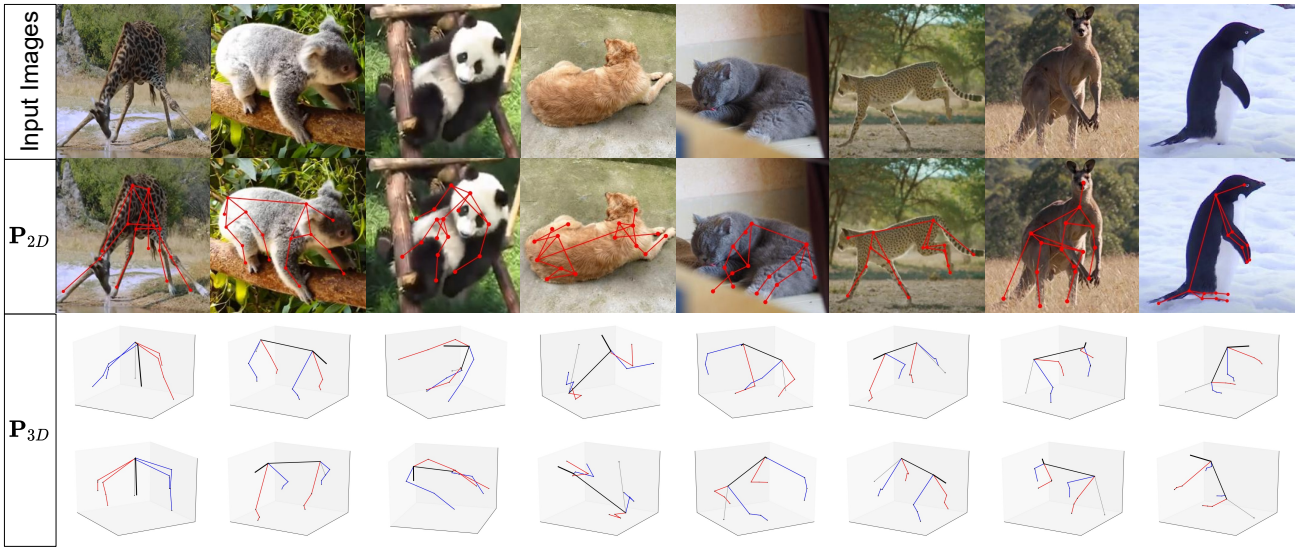

Figure 8. Animal qualitative results on self-collected data. More results can be found in the supplementary material.

2D landmarks to our 2D-to-3D network to retrieve the 3D animal poses and used the reprojection error as the evaluation metric since 3D ground truth annotations are not available. Table 4 shows the error of different animal species. When computing the error, the skeletons are normalized to the range of [-6,6]. Figure 8 shows qualitative results on self-collected animal data, which are complex with very different poses, different limb lengths (only one leg is slightly misaligned in *giraffe* possibly due to perspective shortening), and various degrees of occlusion (the occluded front limb of the black *cat* may actually dip down).

Species	horse	panda	raccoon	cub	bear	alpaca
Error	0.133	0.320	0.277	0.284	0.196	0.244

Table 4. Mean reprojection error for animal estimation. We have 15 horse videos and 1 video for each of the tested animal species.

6. Conclusion

We present a new semi-supervised 3D human/animal pose estimation framework that requires no human-annotated 3D data and enables easy adaptation on encountering unusual poses via synthesizing a massive ground truth 2D-to-3D poses automatically. A large-scale synthetic animal dataset will be released with our unsupervised and 2D-to-3D network. Extensive experiments demonstrate that our method is capable of extreme human/animal pose estimation of different species and achieves the state-of-the-art in 3D human pose estimation. Our future works include: first, the framework can be directly applied to other scenarios, e.g., 3D hand gestures estimation by simply replacing the skeletal model; second, extension to temporal domain in multi-object and partially-visible scenarios.

References

[1] Diogo C Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *CVPR*, 2018. 1

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 1, 2

[3] 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 *ECCV*, 2014. 1, 2

[4] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 36(7), 2013. 1, 2, 7

[5] Shichao Li, Lei Ke, Kevin Pratama, Yu-Wing Tai, Chi-Keung Tang, and Kwang-Ting Cheng. Cascaded deep monocular 3d human pose estimation with evolutionary training data. In *CVPR*, 2020. 1, 2, 3, 4, 7

[6] Bjorn Browatzki and Christian Wallraven. 3fabrec: Fast few-shot face alignment by reconstruction. In *CVPR*, pages 6110–6120, 2020. 1, 2, 3, 4

[7] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, pages 1653–1660, 2014. 2

[8] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *CVPR*, 2015. 2

[9] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016. 2

[10] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. In *CVPR*, 2016. 2

[11] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 2

[12] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018. 2

[13] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *CVPR*, June 2020. 2

[14] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *CVPR*, 2020. 2

[15] Jian Wang, Xiang Long, Yuan Gao, Errui Ding, and Shilei Wen. Graph-pcnn: Two stage human pose estimation with graph pose refinement. *arXiv preprint arXiv:2007.10599*, 2020. 2

[16] Sijin Li and Antoni B Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *ACCV*. Springer, 2014. 2

[17] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation= 2d pose estimation+ matching. In *CVPR*, 2017. 2

[18] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *ICCV*, 2017. 2, 7

[19] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018. 2

[20] Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation from multiple views. In *CVPR*, 2019. 2

[21] Bastian Wandt and Bodo Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *CVPR*, 2019. 2

[22] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *ECCV*, 2020. 2

[23] Xiangyu Xu, Hao Chen, Francesc Moreno-Noguer, Laszlo A. Jeni, and Fernando De la Torre. 3d human shape and pose from a single low-resolution image with self-supervised learning. In *ECCV*, 2020. 2

[24] Zerui Chen, Yan Huang, Hongyuan Yu, Bin Xue, Ke Han, Yiru Guo, and Liang Wang. Towards part-aware monocular 3d human pose estimation: An architecture search approach. In *ECCV*, 2020. 2

[25] Ailing Zeng, Xiao Sun, Fuyang Huang, Minhao Liu, Qiang Xu, and Stephen Lin. Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach. In *ECCV*, 2020. 2

[26] Jogendra Nath Kundu, Siddharth Seth, Varun Jampani, Mugalodi Rakesh, R. Venkatesh Babu, and Anirban Chakraborty. Self-supervised 3d human pose estimation via part guided novel image synthesis. In *CVPR*, 2020. 2

[27] Jingwei Xu, Zhenbo Yu, Bingbing Ni, Jiancheng Yang, Xiaokang Yang, and Wenjun Zhang. Deep kinematics analysis for monocular 3d human pose estimation. In *CVPR*, 2020. 2

[28] Hai Ci, Xiaoxuan Ma, Chunyu Wang, and Yizhou Wang. Locally connected network for monocular 3d human pose estimation. *TPAMI*, 2020. 2

[29] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, volume 2, page 5, 2010. 2, 6, 7

[30] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *CVPR*, 2018. 2

[31] Ben Sapp and Ben Taskar. Modec: Multimodal decomposable models for human pose estimation. In *CVPR*, 2013. 2

[32] 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. *IJCV*, 87(1-2):4, 2010. 2

[33] 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 *3DV*, 2017. 2, 7

[34] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *CVPR*, 2017. 2

[35] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, pages 7297–7306, 2018. 2

[36] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 2

[37] James Charles, Tomas Pfister, Derek Magee, David Hogg, and Andrew Zisserman. Personalizing human video pose estimation. In *CVPR*, 2016. 2

[38] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *NIPS*, 2017. 2

[39] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *CVPR*, 2018. 2

[40] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *CVPR*, 2019. 2

[41] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-identification. In *ECCV*, 2018. 2

[42] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *CVPR*, 2018. 2

[43] Artsiom Sanakoyeu, Vasil Khalidov, Maureen S. McCarthy, Andrea Vedaldi, and Natalia Neverova. Transferring dense pose to proximal animal classes. In *CVPR*, 2020. 2

[44] Siyuan Li, Semih Gunel, Mirela Ostrek, Pavan Ramdya, Pascal Fua, and Helge Rhodin. Deformation-aware unpaired image translation for pose estimation on laboratory animals. In *CVPR*, 2020. 2

[45] Alexander Mathis, Pranav Mamidanna, Kevin M Cury, Taiga Abe, Venkatesh N Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. Technical report, Nature Publishing Group, 2018. 2

[46] Tanmay Nath, Alexander Mathis, An Chi Chen, Amir Patel, Matthias Bethge, and Mackenzie W Mathis. Using deeplabcut for 3d markerless pose estimation across species and behaviors. *Nature protocols*, 14(7):2152–2176, 2019. 2

[47] Muhammad Haris Khan, John McDonagh, Salman Khan, Muhammad Shahabuddin, Aditya Arora, Fahad Shahbaz Khan, Ling Shao, and Georgios Tzimiropoulos. Animalweb: A large-scale hierarchical dataset of annotated animal faces. In *CVPR*, 2020. 3

[48] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 3

[49] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 3

[50] Luca Del Pero, Susanna Ricco, Rahul Sukthankar, and Vittorio Ferrari. Behavior discovery and alignment of articulated object classes from unstructured video. *IJCV*, 121(2):303–325, 2017. 3, 7

[51] Wolfgang Graf, Catherine de Waele, and Pierre Paul Vidal. Functional anatomy of the head-neck movement system of quadrupedal and bipedal mammals. *Journal of Anatomy*, 186(Pt 1):55, 1995. 3

[52] Daniel Schmitt. Mediolateral reaction forces and forelimb anatomy in quadrupedal primates: implications for interpreting locomotor behavior in fossil primates. *Journal of Human Evolution*, 44(1):47–58, 2003. 3

[53] Hamza Khan, Roy Featherstone, Darwin G Caldwell, and Claudio Semini. Bio-inspired knee joint mechanism for a hydraulic quadruped robot. In *ICARA*, 2015. 3

[54] M. Andriluka, U. Iqbal, E. Ensafutdinov, L. Pishchulin, A. Milan, J. Gall, and Schiele B. PoseTrack: A benchmark for human pose estimation and tracking. In *CVPR*, 2018. 3

[55] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 6

[56] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010. 6

[57] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[58] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. In *ICCV*, 2017. 6

[59] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, Oct 2017. 7

[60] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. In *CVPR*, 2018. 7

[61] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *CVPR*, 2019. 7

[62] Saurabh Sharma, Pavan Teja Varigonda, Prashast Bindal, Abhishek Sharma, and Arjun Jain. Monocular 3d human pose estimation by generation and ordinal ranking. In *ICCV*, 2019. 7

[63] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *ICCV*, 2019. 7

[64] 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 *CVPR*, 2018. 7

[65] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Self-supervised learning of 3d human pose using multi-view geometry. In *CVPR*, 2019. 7

[66] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*, 2019. 7

[67] Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *CVPR*, 2019. 7

[68] 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. *TOG*, 36(4):1–14, 2017. 7

[69] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *CVPR*, 2017. 7

[70] 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 *3DV*, 2018. 7

[71] Chenxu Luo, Xiao Chu, and Alan Yuille. Orinet: A fully convolutional network for 3d human pose estimation. *arXiv preprint arXiv:1811.04989*, 2018. 7

[72] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 7