Unsupervised Human Pose Estimation through Transforming Shape Templates

Luca Schmidtke[1], Athanasios Vlontzos[1], Simon Ellershaw[1], Anna Lukens[3],
Tomoki Arichi[2], and Bernhard Kainz[1]

[1]Imperial College London, [2]King's College London, [3]Evelina London Children's Hospital

Abstract

Human pose estimation is a major computer vision problem with applications ranging from augmented reality and video capture to surveillance and movement tracking. In the medical context, the latter may be an important biomarker for neurological impairments in infants. Whilst many methods exist, their application has been limited by the need for well annotated large datasets and the inability to generalize to humans of different shapes and body compositions, e.g. *children and infants. In this paper we present a novel method for learning pose estimators for human adults and infants in an unsupervised fashion. We approach this as a learnable template matching problem facilitated by deep feature extractors. Human-interpretable landmarks are estimated by transforming a template consisting of predefined body parts that are characterized by 2D Gaussian distributions. Enforcing a connectivity prior guides our model to meaningful human shape representations. We demonstrate the effectiveness of our approach on two different datasets including adults and infants. Project page:* infantmotion.github.io

1. Introduction

In today's digitized world, images and videos are an almost endless source of unlabeled, but inherently structured data. Tapping into this reserve of information and knowledge requires the ability to reason in an unsupervised capacity; one of the most compelling and fundamental open problems in machine learning and computer vision.

Self-supervision approaches have shown evidence that they can provide a good supervisory signal for video data [24]. In video recordings an object usually maintains its intrinsic feature distribution but changes its predominantly linear relationships between localized features [23].

In this paper we consider the problem of human pose estimation. Motivated by a wide range of applications including motion capture, visual surveillance and robot control, a continuous effort has been put into generating datasets and models where a manually annotated ground truth pose and key points are available as labels for full supervision. These are currently the most attractive approaches for industrial applications due to their promise of higher accuracy.

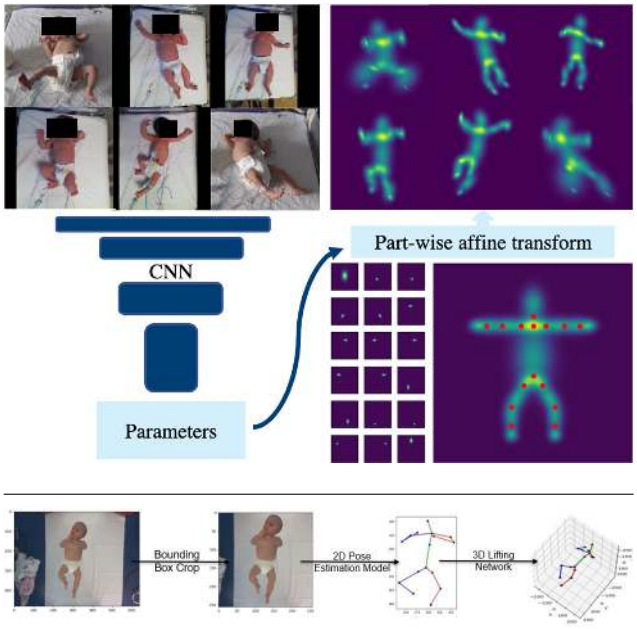

Figure 1. Schematic overview of our approach. Top: We define a part-based human template consisting of 2D Gaussian ellipses and estimate the transformation parameters to estimate the pose of humans with any body composition. Anchor-points are defined between adjacent body parts in order to enforce a connectivity constraint. Bottom: We also evaluate downstream applicability by estimating 3D body poses.

However, ground truth generation is laborious and often limited to a narrow domain, for example, standard poses of healthy adults. In domains with limited demand or special requirements, extensive labeling efforts are often not jus-

tified. Such domains include medical applications, where key point definitions may vary according to diagnostic aims and body shapes might not comply with the learned expectations from a standard training set. Indeed, motion tracking has a variety of applications in medicine, for example to examine the progression of neurological disease and to evaluate treatment success [29], the assessment of injury [5] or for the early diagnosis of impaired neurological development in infants [50, 13]. None of these applications allow for excessive data collection and annotation, often because of a limited number of subjects, restrictions on recording in the clinical environment and economical considerations. Moreover, the direct application of models trained on common benchmark datasets [22, 3, 17] is often challenging.

To tackle this issue, unsupervised and weakly supervised pose estimation methods [23, 34, 30] decompose images into *appearance*, which encodes individual differences such as clothing or body height and *pose*, describing the individual's positioning and configuration of limbs and joints with a canonical latent code. In this context, self-supervision tasks, such as image reconstruction or translation, have been shown to be powerful tools to estimate pose as a *factor of variation* across images instead of relying on strong, manual supervision signals.

We therefore aim at learning the 2D geometry of object categories such as humans and infants with no additional supervision. We exploit the structured information provided by raw videos of continuous pose changes and propose to control inductive bias directly for arbitrary object categories through the manual definition of very simple templates. Thus, we intend to automatically train a neural representation that can predict the 2D pose from a single input image. We show that if such a 2D pose prediction is accurate and compliant with an expected shape prior, these estimations can be extrapolated to 3D poses with a lower error than other existing methods.

We present a method for the unsupervised estimation of 2D keypoints requiring only a simple template and an unannotated video of a single human performing actions in front of a static background to learn a meaningful pose representation. Inspired by previous work [24], where this problem is framed as a *conditional image generation and translation* approach, pose information is utilized to recover a particular frame of a video from any other randomly chosen timepoint. Despite the effective use of self-supervision and representational bottleneck, this approach still requires another prior in the form of unpaired labels and introduces susceptibility to domain shift if these labels come from a different datset. Our model however does not require an additional dataset of unpaired 2D pose examples and relies solely on a simple 2D template consisting of connected body parts modeled as 2D Gaussians. The update of these Gaussians can be learned as affine transformations. Even though images are 2D representations of 3D information, affine transformations allow to model all possible projected configurations of body parts. Motivated by part-based approaches such as [64] we introduce anchor-points in order to enforce connectivity between body parts and regularize model training and prediction.

In summary, we make the following contributions:

- We introduce a conceptually simple but effective method to learn 2D human-interpretable keypoints based on transforming a single manually defined 2D template.
- Our proposed approach is capable of performing 2D human pose estimation without any additional need for labeled data, either paired or unpaired.
- We demonstrate the high adaptability of our approach by evaluating it on benchmark data and in the wild on a challenging infant pose estimation dataset.

2. Related Work

We consider the problem of predicting the 2D pose of an object from a single 2D RGB image as a pose recognition task. We structure related methods according to **full supervision**, where dense manual annotations are paired to each frame in a dataset, and **weak supervision or no supervision**, when only partial or no annotations are available or when models are transferred to a new domain. The question when a method can be referred to as weakly supervised or unsupervised remains debatable, which is why we do not make this distinction explicit.

Inductive bias and prior knowledge are also important in this context. Very often latent manifold distributions that are known a priori are used as an initialization step. This can include models that have been learned from other data or any other type of supervision or empirical priors. Similar to recent attempts in this field, *e.g.* [23], our method is unsupervised but it uses a very simple empirical prior, which we hypothesize leads to better results.

Fully supervised models rely on carefully annotated data, which is available for narrow fields of applications, *e.g.* MS COCO keypoints [32], MPII Human Pose database [3], Human3.6M [22] or LSP [17]. Methods utilizing these datasets are usually trained without additional priors due to the abundance of direct labels. Pictorial structures [12, 4, 39, 41, 43, 45, 62] have been used to describe poses and CNNs have shown evidence to be powerful estimators for keypoints [55] and their uncertainties [54]. Confidence heatmaps are popular for scenes where a single pose needs to be estimated [6, 7, 9, 37, 40, 53, 58] or multiple poses at once [8, 21]. Our framework does not use existing image annotation to learn a pose prediction model.

Weakly supervised and unsupervised methods including self-supervised approaches have grown in popularity due to their efficiency in dealing with limited ground truth data. This becomes especially important when there are no large publicly available datasets available for the target domain. Jakab et al. [23] propose to learn a pose representation via conditional image translation. By leveraging the extracted pose representation and given another frame containing the same person in a different pose they task the model to reconstruct the original image. However, a pixel-wise reconstruction loss does not encourage meaningful, low-dimensional representations. To avoid the model from simply encoding pixel information similar to an autoencoder, a representational bottleneck in the form of k tuples of (x, y) coordinates is introduced. This enables unsupervised pose landmark detection provided that the background remains static. However, the extracted landmarks follow no common convention and are difficult or impossible to interpret. To tackle this, [24] further propose the introduction of a fully differentiable image-based representation, resembling a human skeleton. Adversarial training with unpaired pose labels is required to make the model converge towards human-interpretable outputs.

Zhang et al. [63] introduce equivariance and invariance constraints under an autoencoder-based formulation, while Lorenz et al. [34] expand this approach by introducing disentanglement between object pose and appearance. Kundu et. al. proposed to use an energy-based optimization approach combined with a part-based shape template to estimate 3D poses in images with varying backgrounds [30]. The method produces impressive results but still relies on a set of unpaired real 3D pose labels.

Earlier methods learn to predict dense 3D human meshes from sparse 2D keypoint annotations, *e.g.* [25], by using a parametric human mesh model [33] and regularization by adversarial learning from motion capture coordinates. Related approaches have been proposed in [14, 15, 16, 36, 46, 56, 57, 61].

Others propose methods to match pairs of images of an object, but sacrifice geometric invariants such as keypoints to achieve this [26, 44, 47]. Sparse and dense landmarks are introduced by [49, 52, 51] in the unsupervised context. Synthetic views of 3D models like in [49, 20] are a reasonable workaround which our method does not use.

Other noteworthy approaches learn a dense deformation field, *e.g.* for faces [47, 59]. In contrast to their methods and similar to [24] we predict semantically meaningful keypoints, however, our points can be freely defined and easily changed through a simple template. Thus, the quality of our landmarks is higher per definition of our approach and adaptable for any given application.

Human pose estimation in medicine: Medical applications pose a larger challenge for pose estimation algorithms. For infants, existing learned pose estimation models suffer from domain shift, thus, focused models have been developed. Pose Estimation in videos of infants has been proposed as an early diagnosis tool of diseases related to impaired neurological development affecting the sensorimotor system [50, 13]. In previous work, marker-based approaches relying on optical [2] or electromagnetic [27] tracking have been utilized to track motion in infants. However, these methods rely on clinical specialists, expertise, often costly equipment and a substantial amount of manual preparation and calibration. Different marker-less approaches have been proposed based on optical flow [48] or particle matching [42]. With the advent of compact and cheap cameras with integrated depth sensors, several more recent works combine images and depth information with random fern classifiers [20], a deformable parts model [28] or by employing a shape model [38, 19].

For medical applications 3D pose prediction holds additional value. The most common approach to 3D pose estimation is full supervision as detailed above. One of the earliest approaches outlining a supervised deep learning approach to the task was proposed by Li et al. [31] with a network analogous to [55]. In [11], the authors proposed splitting the human pose estimation task into two parts. The first is a generic 2D pose estimation model using CNNs, the second is a non-parametric nearest neighbor model that paired the estimated 2D pose to the closest 2D pose from a dictionary of paired 2D and 3D poses. Martinez et al. [35] took this a step further, replacing the 3D dictionary lookup model with a deep learning 3D lifting network that took the estimated 2D pose vectors as an input and produced 3D 'lifted' outputs. The use of a differentiable soft argmax function [10] allows end-to-end training of a fully differentiable model and an L2 loss function can be used to directly regress 3D keypoint locations.

3. Approach

In this section we present our proposed method. We draw inspiration from [24] and formulate a self-supervision task by tasking our model to reconstruct an initial frame $\mathbf{f}_t$ conditioned on an estimated low-dimensional pose representation $\tilde{p}_t$. On a high level, our method is composed of two modules. The first network φ, given the initial frame $\mathbf{f}_t$ and a template $\mathbf{T}$ consisting of 2D Gaussian heatmaps of human body parts, extracts an estimate $\tilde{p}_t$ of the true pose p_t represented by a spatial arrangement of different body parts. The second module is an encoder-decoder network ϕ which receives as input the previously extracted pose together with a frame $\mathbf{f}_{t+k}$ containing the same person in a different pose

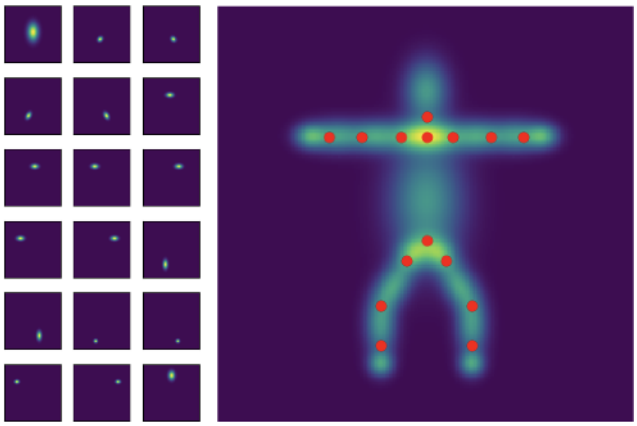

Figure 2. Left: Body part heatmaps; Right: combined template with defined anchor points plotted in red.

p_{t+k}, and tasked to reconstruct the original frame f_t.

$$\tilde{\mathbf{f}}_t = \phi(\mathbf{f}_{t+k}, \tilde{p}_t) \quad (1)$$

$$\tilde{p}_t = \varphi(\mathbf{f}_t, \mathbf{T}) \quad (2)$$

The pose representation $\tilde{p}_t$ is found by optimizing the model to encode all necessary information to recover the original pose whilst enforcing representational constraints to avoid the encoding of low-level image features.

3.1. Template and Anchor-points

Inspired by [30] and [64], we design a human template as shown in Figure 2 that consists of 18 body parts in an effort to represent human anatomy (forearm, torso, head, etc.). Instead of transforming this template according to previously found keypoints in the image, we directly estimate the transformations and make use of the template as a strong prior. Each part is characterized by its central position (x_0, y_0) surrounded by a 2D Gaussian with variance (Var_x, Var_y). We pre-define the mean and variance of each Gaussian for all body parts in an effort to represent the length and width of the individual parts. Hence, we are able to model a canonical T-pose as shown in Figure 2. This pose was chosen because it roughly represents the mean of possible pose configurations. Each representation of a body part $\mathbf{l} \in \mathbb{R}^{h \times w}$ is saved in an individual channel resulting in a tensor of shape (B, K, H, W) where B is the batch size, K the number of body parts and H, W are width and height of the input image.

For each body part $\mathbf{l}_j$ in our template $\mathbf{T}$ we define N_j (up to three depending on which part) anchor points $\mathbf{a}_i^j \in \mathbb{R}^2$ in image space, with $j = 1, ..., K$ and $i = 1, ..., N_j$. Our template design and the resulting anchor points coincide with the most commonly used landmark definitions for human pose estimation. However, the detected landmarks can be easily changed by simply choosing different points on each body part.

3.2. Pose Extractor

The network $\varphi : \mathbb{R}^{3 \times h \times w} \rightarrow \mathbb{R}^{k \times 3 \times 3}$ is implemented as a fully-convolutional neural network followed by a fully-connected layer. An image $\mathbf{f}_t$ is passed through a series of down-sampling convolutional layers and mapped to a pose representation described by the parameters of affine transformations $\Theta_k \in \mathbb{R}^{3x3}$ for each k-th body part from the original template:

$$\Theta = \begin{bmatrix} \theta_{1,1} & \theta_{1,2} & t_x \\ \theta_{2,1} & \theta_{2,2} & t_y \\ 0 & 0 & 1 \end{bmatrix}, \quad (3)$$

where $t_{x,y}$ represent the translations across the two dimensions and θ correspond to rotation, shear, and scale. Each part of the template is then warped by its corresponding transform, resulting in $\mathbf{T}' = (\vartheta_0(\mathbf{l}_0), ..., \vartheta_k(\mathbf{l}_k))$, where ϑ_i is the corresponding transform for each Θ_i.

3.3. Image Translation Module

The network $\phi : \mathbb{R}^{3+k \times h \times w} \rightarrow \mathbb{R}^{3 \times h \times w}$ receives a different frame $\mathbf{f}_{t+k}$ along with the transformed template $\mathbf{T}'$ concatenated along the channel dimension and outputs another image $\tilde{\mathbf{f}}_t$ following an encoder-decoder pathway. Intuitively, this network learns to implicitly disentangle content (*i.e.*, person identity) and pose in an image and transforms the content according to the conditioning pose.

3.4. Training Objectives

Our training objective is split into three parts: A reconstruction loss is associated with the image translation module while a boundary and an anchor loss guide the training of the pose extractor network.

Reconstruction Loss The main loss component in our method is the reconstruction objective between the original frame $\mathbf{f}_t$ and the output of the image translation network ϕ. Similar to [24], we found it to be beneficial to use a perceptual loss with a pretrained VGG network in order to stabilize training:

$$\mathcal{L}_{recon} = ||\psi(\tilde{\mathbf{f}}_t) - \psi(\mathbf{f_t})||_1^1, \quad (4)$$

where $\psi(\mathbf{f})$ are feature vectors extracted from a frame by the VGG network and $\tilde{\mathbf{f}}_t$ the output of the image translation network.

Anchor-point Loss Each anchor point is being transformed by the corresponding body part transform: $\tilde{\mathbf{a}}_i^j = \Theta_j \mathbf{a}_i^j$. In order to enforce connectivity between body parts,

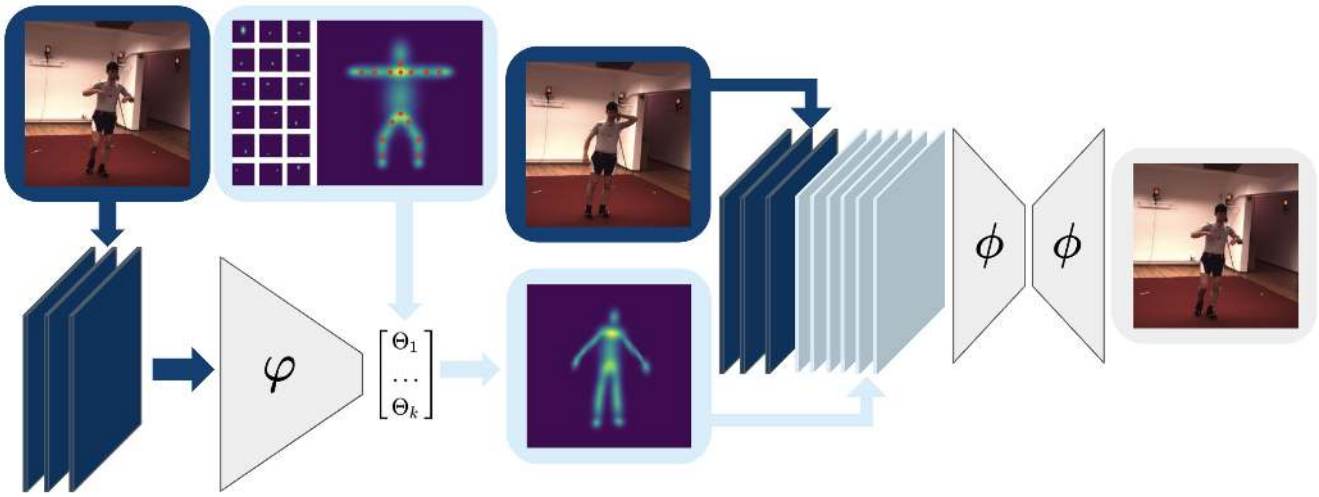

Figure 3. Schematic overview of our approach: We define a part-based human puppet template and predict the transformation parameters to estimate the pose for an input frame.

we define a set $A = \{(\tilde{\mathbf{a}}_j^l, \tilde{\mathbf{a}}_k^m) | l \neq m\}$ containing all pairs of transformed anchor points which we require to have the same position. Lastly we compute the mean squared L2 distance of the two points over all tuples:

$$\mathcal{L}_{\text{anchor}} = \frac{1}{M} \sum_{\tilde{\mathbf{a}}_j^l, \tilde{\mathbf{a}}_k^m \in A} ||\tilde{\mathbf{a}}_j^l - \tilde{\mathbf{a}}_k^m||_2^2, \tag{5}$$

where $M =$ number of anchor point pairs.

Boundary Loss Additionally, we found it beneficial for convergence and to prevent the network from outputting transformation parameters leading to a translation of the template shapes out of the image boundaries to enforce the anchor points to be contained within the image:

$$\mathcal{L}_{bx} = \begin{cases} |a_{i,x}^j|, & \text{if } |a_{i,x}^j| > B \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

where $a_{i,x}^j$ is the x-coordinate of the anchor point in pixel space and B corresponds to the size of the (quadratic) image. The loss $\mathcal{L}_{by}$ for the y-component is analogous.

Combined Loss Formulation In summary our training objective is expressed as

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda_1 \mathcal{L}_{\text{anchor}} + \lambda_2 (\mathcal{L}_{bx} + \mathcal{L}_{by}). \tag{7}$$

Training Our whole model including all subnetworks is trained in an end-to-tend manner. For sampling pairs, we define a course grid and corresponding bounding boxes based on the supplied masks and select frames from the same bounding box to ensure a static background. Note, that the hyper-parameters λ_1, λ_2 are empirically tuned.

4. Experiments

We evaluate our approach on two different datasets, Human3.6m and an infant dataset from clinical practice. We compare our method to results with various degrees of supervision, as reported in recent literature. Additionally, we performed an ablation study to demonstrate the effectiveness of our proposed anchor and boundary losses.

Human3.6m is an industry standard dataset for human pose estimation [22] containing 3.6 million images with corresponding 3D and 2D landmarks and captured with different actors in a controlled studio environment with a static background.

Infant Dataset is a clinical dataset containing videos of 24 infants captured in their cod in the hospital. The dataset was recorded and curated by the authors and their clinical collaborators using an image labelling tool [1]. All subjects were recruited from consenting parents while the study and its use has been cleared by the appropriate ethics committees. We train our method on 20 subjects and evaluate on four, resulting in a total of 290k unlabelled training and 471 manually labelled test images.

Moving Infants In RGB-D (MINI-RGBD) consists of 12 synthetically rendered movement sequences of infants with different body shapes and backgrounds [18]. The dataset was designed to cover challenging and diverse movements and overall contains 12k images together with various label information including landmarks and masks. We introduce this dataset to demonstrate the issue of domain shift when using these labels as an additional prior.

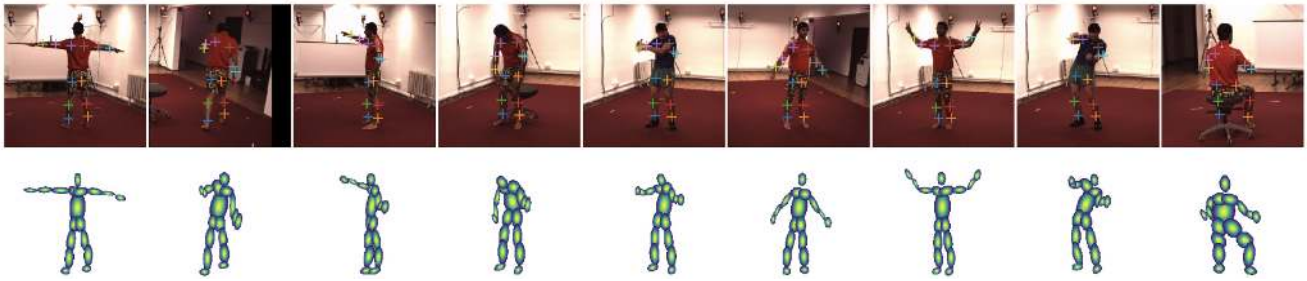

Figure 4. Results of our model on random frames from subjects 9 and 11 in Human 3.6m. Top row: input images with detected keypoints. Bottom row: corresponding deformed shape templates

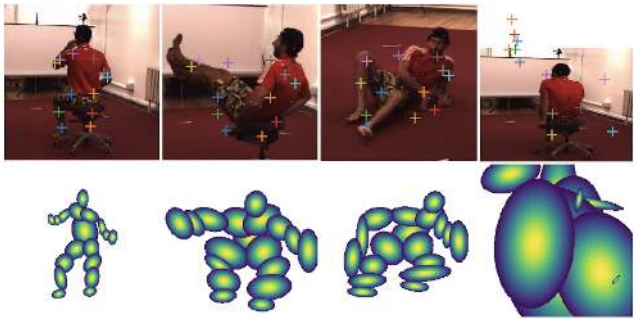

Figure 5. Example for limitations of our model. While sitting is a very difficult pose, facing away from the camera is the most challenging situation. We observe the full loss of coherent landmarks in some of these cases. Moreover, the model switches left and right when a person is facing away from the camera.

4.1. Evaluation Procedure

Human 3.6m In order to make our results comparable to recent literature with regards to the Human3.6m dataset, we adopt the evaluation strategy from [34, 63]. The model is trained on subjects 1,5,6,7,8 and tested on subject 9. We restrict the performed actions to {direction, discussion, posing, waiting, greeting, walking} during testing in order to ensure mostly upright poses, resulting in a total of 80k images for testing. 16 landmarks are predicted and compared with the ground truth. More results can be found in the supplemental material. The comparison methods make use of the supplied estimated rough person masks and substract the background. We did not perform this extra preprocessing step.

Many proposed unsupervised or weakly supervised methods require an additional post-processing step and direct supervision via ground truth 2D keypoints. Most use a linear regressor, in order to output human-interpretable landmarks. Our approach returns interpretable landmarks in form of anchorpoints *by design*. The latter also coincide with the most commonly used conventions, since they are placed around the joints.

	BBox	SPP	UL	T
Lorenz [34]	✓	✓	✗	✗
Zhang [63]	✓	✓	✗	✗
Jakab [24]	✓	✗	✓	✗
Ours	✓	✗	✗	✓

Table 1. Comparing different sources of labels used for different methods. **BBox**: bounding box centered around the person, **SPP**: supervised post-processing. A linear regressor is applied to map discovered landmarks to human-interpretable locations. **UL**: Use of unpaired manually annotated poses as a prior. **T**: the manual design of a single template including individual body parts and anchorpoints

Infant Dataset As a baseline, we fine-tune a fully supervised ResNet based network [60] pretrained on both ImageNet and adult poses from MPII Human Pose [3]. For further comparison, we implemented and trained the method proposed by Jakab et al. [24] on our clinical infant images combined with unpaired poses from the synthetic MINI-RGBD infant dataset [18] in order to demonstrate the method's susceptibility to domain shift. 12 landmarks are predicted and compared with the ground truth.

For a fair comparison of the results, we make this explicit by adding *supervised post-processing* in parenthesis for all results where this step was included. We also compiled Table 1 to clearly indicate the different sources of labels that are used by the different methods.

4.2. Results

Human 3.6m For adult pose estimation, we summarize our results in Table 2. As expected, none of the self-supervised methods perform as well as the supervised baseline due to the lack of labels. All prior work makes use of paired or unpaired manual annotations in some capacity. Despite the complete lack of such annotations, our method performs competitively on the same task while only requiring the template. Our model is also able to predict landmarks for difficult poses such as sitting on a chair, as can be seen in Figure 4. Errors are largest when self-occlusion occurs. In the most severe cases, when a person is facing

Figure 6. Results of our model on random frames from the test subjects of our in-house dataset. Anonymisation was applied to maintain patient privacy. Top row: input images with detected keypoints. Bottom row: corresponding deformed shape templates

H36M	all	wait	pose	greet	direct	discuss	walk
fully supervised baseline							
Newell [37]	2.16	1.88	1.92	2.15	1.62	1.88	2.21
self-supervised + supervised post-processing							
Thewlis [52]	7.51	7.54	8.56	7.26	6.47	7.93	5.40
Zhang [63]	4.14	5.01	4.61	4.76	4.45	4.91	4.61
Lorenz [34]	2.79	–	–	–	–	–	–
self-supervised (unpaired labels)							
Jakab [24]	2.73	2.66	2.27	2.73	2.35	2.35	4.00
self-supervised (template, no labels)							
Ours	3.31	3.51	3.28	3.50	3.03	2.97	3.55

Table 2. Comparison with state-of-the-art methods for human landmark detection on the Simplified Human3.6M dataset; %-MSE normalized by image size is reported on a per action basis. Note that our method does not require any annotations at all while results are en-par with the state-of-the-art unsupervised approaches utilizing unpaired labels or post-processing.

Infants	all	hips	knees	feet	shoulders	hands	params
fully supervised (fine-tuned) baseline							
Xiao [60]	1.74	2.39	1.50	1.47	1.76	1.59	34.0 M
self-supervised (unpaired labels)							
Jakab [24]	8.98	6.89	8.18	13.15	5.33	11.36	8.6 M
self-supervised (template, no labels)							
Ours	4.86	3.79	4.60	5.53	3.19	7.21	7.8 M

Table 3. Comparison with state-of-the-art methods for the Infant dataset; %-MSE normalized by image size is reported on a per body-part basis. Our method outperforms prior unsupervised approaches for this task because it is not influenced by annotation domain shift.

away from the camera and both arms and hands are covered by the body, estimating these landmarks becomes extremely difficult without further supervision in the form of manually annotated examples or additional views captured by a second camera.

Infant Pose Estimation The results for infant pose estimation are summarized in Table 3 and 5. Figure 6 displays predictions of our model on infants. Our implemented version of [24] performs worse despite the access to 11,000 unpaired landmark annotations. We attribute this to domain shift. Since the model relies on adversarial training on these labels, the performance will drop if the latter are not covering a diverse enough range of possible poses. In fact, the authors demonstrate a drop in performance in their own experiments when using labels from a different dataset.

Our model is capable of predicting consistent and interpretable landmarks from images of infants with different body shapes and in different poses. Again, the largest errors are introduced by self-occlusion, especially when arms are positioned in front of the chest and shoulders. This is consistent with our observation that the landmark detection works best on the legs, which are most of the time not positioned above or below other body parts.

Ablation study In order to verify the individual contribu-

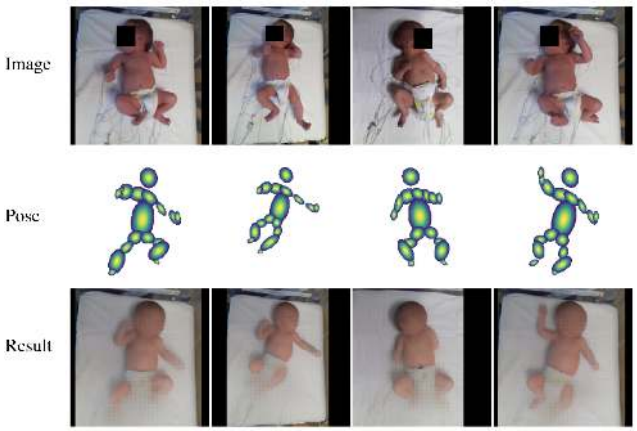

Figure 7. Pose conditioning. Top row: Input image; Middle row: New Conditioning Pose; Bottom row: Resulting image. We observe that our image translation sub-module is able to disentangle pose and content and produce photo-realistic images with the conditioning poses.

tion of our loss components, we partially train several models with different loss configurations and present the results in Table 4.

all	hips	knees	feet	shoulders	hands
anchor and boundary loss (ours)					
6.54	3.56	5.19	8.90	4.70	7.95
with anchor, no boundary loss					
65.99	64.65	71.70	76.33	59.00	62.59
with boundary, no anchor loss					
12.42	13.84	13.17	7.29	12.94	8.76
no anchor, no boundary loss (model diverges)					
446.47	433.64	385.77	662.41	502.69	276.49

Table 4. Mean joint distance in % of image size (values>100 are outside of image). All models trained for 10 epochs.

Image Translation Since our decoder network is trained to synthesize images conditioned on the pose input, we are able to perform image editing by combining a frame $\mathbf{f}$ with a different set of poses p. The images displayed in Figure 7 illustrate the principle and demonstrate disentanglement between pose and appearance. Moreover, the image translation modules' ability to produce images given a conditioning pose could serve as a tool in the analysis for indicators of impaired neurological development. This, however, remains a topic of future work.

Limitations Our method is limited by occlusions and partial views of the body and is currently not able to distinguish back-facing from front-facing poses. If not accounted for, this can result in large errors due to the switching of joints with their left/right counterpart. The problem only occurs for adult pose estimation, as infants are lying on their backs during a recording and are physically not able to turn around. We believe that the inclusion of a directional vector could alleviate the observed limitation.

5. Downstream application study

To study the downstream effects of 2D pose prediction accuracy in a realistic setting we integrate our method into a setup as it is currently used for clinical evaluation of neonatal movement quality. Physiological movement quality assessment requires 3D coordinates, thus a fully supervised 3D lifting network is applied after 2D pose estimation. The lifting model is adapted from the work of Martinez et al. [35]. This model was first pre-trained on the MPI_INF_3DHP adult dataset [36] before fine-tuning on the infant MINI-RGBD dataset [18]. Figure 8 compares the the qualitative differences when results from previous 2D pose estimation methods [24] are used or ours. Quantitatively this can be captured by evaluating the joint position error in mm. These results are summarized in Table 5.

Synthetic Infants	all	hips	knees	feet	shoulders	hands
self-supervised (unpaired labels)						
Jakab [24]	59.7	10.1	58.4	73.3	58.0	101.5
self-supervised (template, no labels)						
Ours	44.7	8.0	57.5	80.9	35.6	78.7

Table 5. 3D lifting results. The error is reported as the distance between predicted and ground truth 3D landmarks in mm.

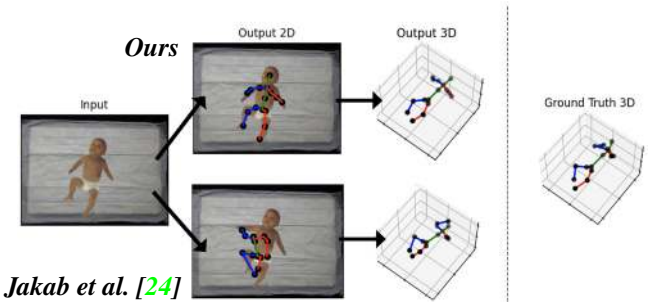

Figure 8. Visual results from the lifting network on a synthetic infant image. Top: ours, bottom: Jakab et al. [24]. The accuracy of the predicted 2D keypoints has a noticeable impact on the final 3D predictions.

6. Conclusion

We presented a novel approach to unsupervised human pose estimation by transforming a part-based shape template. Given a video of a person moving in front of a static background, we are able to predict human interpretable keypoints without requiring any paired or unpaired labels. We have exhibited our method's performance on two different datasets and achieved similar or better results in unsupervised human pose estimation while only requiring a simple canonical 2D pose template instead of numerous manual labels. Moreover, our method is able to adapt to humans or infants of different shapes and sizes and alternative keypoints can be defined easily by choosing different locations on each body part. We also demonstrate an increase in performance when using our method's extracted poses in downstream tasks like 3D pose extrapolation, which is of significance for both industrial and medical applications. Although some manual work is required to define the template, this could be easily realized in a simple GUI, where the user would be able to draw body parts and define anchoras well as keypoints within minutes.

Acknowledgements: supported by EPSRC EP/S013687/1.

References

[1] Labelbox. *Online, Available: https://labelbox.com*, 2021.

[2] Manu Airaksinen, Okko Räsänen, Elina Ilén, Taru Häyrinen, Anna Kivi, Viviana Marchi, Anastasia Gallen, Sonja Blom, Anni Varhe, Nico Kaartinen, Leena Haataja, and Sampsa Vanhatalo. Automatic posture and movement tracking of infants with wearable movement sensors. *Scientific Reports*, 2020.

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[4] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1014–1021. IEEE, 2009.

[5] TS Bae, K Choi, and M Mun. Level walking and stair climbing gait in above-knee amputees. *Journal of medical engineering & technology*, 33(2):130–135, 2009.

[6] Vasileios Belagiannis and Andrew Zisserman. Recurrent human pose estimation. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 468–475. IEEE, 2017.

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

[8] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

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

[10] Olivier Chapelle and Mingrui Wu. Gradient descent optimization of smoothed information retrieval metrics. *Information retrieval*, 13(3):216–235, 2010.

[11] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation= 2d pose estimation+ matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7035–7043, 2017.

[12] Pedro F Felzenszwalb and Daniel P Huttenlocher. Pictorial structures for object recognition. *International journal of computer vision*, 61(1):55–79, 2005.

[13] Fabrizio Ferrari, Giovanni Cioni, Christa Einspieler, Maria Federica Roversi, Arend F Bos, Paola Bruna Paolicelli, Andrea Ranzi, and Heinz R. F. Prechtl. Cramped synchronized general movements in preterm infants as an early marker for cerebral palsy. *Archives of pediatrics & adolescent medicine*, 156 5:460–7, 2002.

[14] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1155–1164, 2019.

[15] Zhenglin Geng, Chen Cao, and Sergey Tulyakov. 3d guided fine-grained face manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9821–9830, 2019.

[16] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn, and Thomas Vetter. Morphable face models-an open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 75–82. IEEE, 2018.

[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] Nikolas Hesse, Christoph Bodensteiner, Michael Arens, Ulrich G Hofmann, Raphael Weinberger, and A Sebastian Schroeder. Computer vision for medical infant motion analysis: State of the art and rgb-d data set. In *European Conference on Computer Vision*, pages 32–49. Springer, 2018.

[19] Nikolas Hesse, Sergi Pujades, Javier Romero, Michael J. Black, Christoph Bodensteiner, Michael Arens, Ulrich G. Hofmann, Uta Tacke, Mijna Hadders-Algra, Raphael Weinberger, Wolfgang Müller-Felber, and A. Sebastian Schroeder. Learning an infant body model from RGB-D data for accurate full body motion analysis. 2018.

[20] N. Hesse, G. Stachowiak, T. Breuer, and M. Arens. Estimating body pose of infants in depth images using random ferns. *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2015.

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

[22] 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, jul 2014.

[23] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 4016–4027. Curran Associates, Inc., 2018.

[24] Tomas Jakab, A. Gupta, Hakan Bilen, and A. Vedaldi. Self-supervised learning of interpretable keypoints from unlabelled videos. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8784–8794, 2020.

[25] 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.

[26] Angjoo Kanazawa, David W Jacobs, and Manmohan Chandraker. Warpnet: Weakly supervised matching for single-view reconstruction. In *Proceedings of the IEEE Conference*

on Computer Vision and Pattern Recognition, pages 3253–3261, 2016.

[27] Dominik Karch, Keun-Sun Kim, Katarzyna Wochner, Joachim Pietz, Hartmut Dickhaus, and Heike Philippi. Quantification of the segmental kinematics of spontaneous infant movements. *Journal of biomechanics*, 41(13), September 2008.

[28] Muhammad Hassan Khan, Manuel Schneider, Muhammad Shahid Farid, and Marcin Grzegorzek. Detection of infantile movement disorders in video data using deformable part-based model. *Sensors*, 18, 2018.

[29] Peter Kontschieder, Jonas F Dorn, Cecily Morrison, Robert Corish, Darko Zikic, Abigail Sellen, Marcus D'Souza, Christian P Kamm, Jessica Burggraaff, Prejaas Tewarie, et al. Quantifying progression of multiple sclerosis via classification of depth videos. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 429–437. Springer, 2014.

[30] Jogendra Nath Kundu, S. Seth, V. Jampani, M. Rakesh, R. Babu, and Anirban Chakraborty. Self-supervised 3d human pose estimation via part guided novel image synthesis. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6151–6161, 2020.

[31] Sijin Li and Antoni B Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *Asian Conference on Computer Vision*, pages 332–347. Springer, 2014.

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, editor="Fleet David Zitnick, C. Lawrence", Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755. Springer International Publishing, 2014.

[33] 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):1–16, 2015.

[34] Dominik Lorenz, Leonard Bereska, Timo Milbich, and B. Ommer. Unsupervised part-based disentangling of object shape and appearance. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10947–10956, 2019.

[35] 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.

[36] 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.

[37] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 483–499, Cham, 2016. Springer International Publishing.

[38] Mikkel Damgaard Olsen, Anna Herskind, Jens Bo Nielsen, and Rasmus R. Paulsen. Model-based motion tracking of infants. 2014.

[39] Wanli Ouyang, Xiao Chu, and Xiaogang Wang. Multi-source deep learning for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2329–2336, 2014.

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

[41] Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. Poselet conditioned pictorial structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–595, 2013.

[42] Hodjat Rahmati, Ralf Dragon, Ole Morten Aamo, Lars Adde, Øyvind Stavdahl, and Luc Van Gool. Weakly supervised motion segmentation with particle matching. *Computer Vision and Image Understanding*, 2015.

[43] Varun Ramakrishna, Daniel Munoz, Martial Hebert, James Andrew Bagnell, and Yaser Sheikh. Pose machines: Articulated pose estimation via inference machines. In *European Conference on Computer Vision*, pages 33–47. Springer, 2014.

[44] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6148–6157, 2017.

[45] Benjamin Sapp, Chris Jordan, and Ben Taskar. Adaptive pose priors for pictorial structures. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 422–429. IEEE, 2010.

[46] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6296–6305, 2018.

[47] Zhixin Shu, Mihir Sahasrabudhe, Riza Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *Proceedings of the European conference on computer vision (ECCV)*, pages 650–665, 2018.

[48] Ayelet Stahl, Christian Schellewald, Oyvind Stavdahl, Ole Morten Aamo, Lars Adde, and Harald Kirkerød. An optical flow-based method to predict infantile cerebral palsy. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 20:605–614, 2012.

[49] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker, and Rudolph Triebel. Implicit 3d orientation learning for 6d object detection from rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 699–715, 2018.

[50] Philip Teitelbaum, Osnat Teitelbaum, J Nye, Joshua B. Fryman, and Ralph G. Maurer. Movement analysis in infancy may be useful for early diagnosis of autism. *Proceedings of the National Academy of Sciences of the United States of America*, 95 23:13982–7, 1998.

[51] James Thewlis, Samuel Albanie, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of landmarks by descriptor vector exchange. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6361–6371, 2019.

[52] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling. In *Advances in neural information processing systems*, pages 844–855, 2017.

[53] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 648–656, 2015.

[54] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1799–1807. Curran Associates, Inc., 2014.

[55] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014.

[56] Hsiao-Yu Fish Tung, Adam W Harley, William Seto, and Katerina Fragkiadaki. Adversarial inverse graphics networks: Learning 2d-to-3d lifting and image-to-image translation from unpaired supervision. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4364–4372. IEEE, 2017.

[57] Mengjiao Wang, Zhixin Shu, Shiyang Cheng, Yannis Panagakis, Dimitris Samaras, and Stefanos Zafeiriou. An adversarial neuro-tensorial approach for learning disentangled representations. *International Journal of Computer Vision*, 127(6-7):743–762, 2019.

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

[59] Olivia Wiles, A Koepke, and Andrew Zisserman. Self-supervised learning of a facial attribute embedding from video. *arXiv preprint arXiv:1808.06882*, 2018.

[60] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *European Conference on Computer Vision (ECCV)*, 2018.

[61] 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.

[62] Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR 2011*, pages 1385–1392. IEEE, 2011.

[63] Y. Zhang, Yijie Guo, Y. Jin, Yijun Luo, Zhiyuan He, and H. Lee. Unsupervised discovery of object landmarks as structural representations. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2694–2703, 2018.

[64] Silvia Zuffi and Michael J. Black. The stitched puppet: A graphical model of 3D human shape and pose. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2015)*, pages 3537–3546, June 2015.