Heuristic Weakly Supervised 3D Human Pose Estimation in Novel Contexts without Any 3D Pose Ground Truth

Shuangjun Liu, Xiaofei Huang, Nihang Fu, and Sarah Ostadabbas

Abstract—Monocular 3D human pose estimation from a single RGB image has received a lot attentions in the past few year. Pose inference models with competitive performance however require supervision with 3D pose ground truth data or at least known pose priors in their target domain. Yet, these data requirements in many real-world applications with data collection constraints may not be achievable. In this paper, we present a heuristic weakly supervised solution, called HW-HuP to estimate 3D human pose in contexts that no ground truth 3D data is accessible, even for fine-tuning. HW-HuP learns partial pose priors from public 3D human pose datasets and uses easy-to-access observations from the target domain to iteratively estimate 3D human pose and shape in an optimization and regression hybrid cycle. In our design, depth data as an auxiliary information is employed as weak supervision during training, yet it is not needed for the inference. We evaluate HW-HuP performance quantitatively on datasets of both in-bed human and infant poses, where no ground truth 3D pose is provided neither any target prior. We also test HW-HuP performance quantitatively on a publicly available motion capture dataset against the 3D ground truth. HW-HuP is also able to be extended to other input modalities for pose estimation tasks especially under adverse vision conditions, such as occlusion or full darkness. On the Human3.6M benchmark, HW-HuP shows 104.1mm in MPJPE and 50.4mm in PA MPJPE, comparable to the existing state-of-the-art approaches that benefit from full 3D pose supervision.

Index Terms—3D Human Pose Estimation, Adverse Vision Conditions, Data-efficient Machine Learning, Data Scarcity, Transfer Learning, Weak Supervision.

1 INTRODUCTION

As the applications of machine learning (ML) expand into our everyday life, the data scarcity issue poses a seemingly insurmountable hurdle for many scientific, industrial, and healthcare applications, where gathering or labeling data can be prohibitively expensive due to sampling cost or strong privacy laws. Data-efficient ML approaches therefore have been trying to exploit structural knowledge in the problem in order to constrain them to a point where the model is simple enough to be correctly trained with the available data [1], [2]. One way is to train an initial inference model using plentiful data from a similar problem, then transfer the learning (e.g. via fine-tuning) to the target problem [3], [4]. Data-efficient ML however often requires some sample data (although limited) from the target domain or at least a series of shared attributes in the zero-shot learning case [5]. But, what if not even a single annotated sample is available in a target application due to data collection difficulties or context constraints?

Let’s look at the problem of 3D human pose estimation in applications in the “wild”, where no ground truth 3D pose data are accessible, neither any known priors. This paper presents an effective transfer learning approach based on a heuristic weakly supervised framework to enable the estimation of 3D human poses under such constraints. Our solution uses easy-to-access 3D proxy data for weak supervision as well as heuristic rule extraction to train a robust 3D human pose estimation model. Our work in particular focuses on applications that the environmental or person-specific constraints prevent the use of professional motion capture (MoCap) devices to collect even a small 3D pose sample set in the target domain. These contexts could be related to patient monitoring in a hospital room [6], [7] or baby monitoring in their crib [8]. Fig. 1 demonstrates that how recent high performance state-of-the-art 3D pose estimation models when applied for the in-bed patient monitoring—an application critically in need for automation especially during the recent pandemic—fail to yield qualitatively correct results. To deal with the difficulties of gathering MoCap data in practical applications, depth sensing using inexpensive
off-the-shelf cameras has been employed in many human pose estimation studies, which adds an extra dimension to the otherwise 2D RGB data [9]. However, RGBD is not 3D and popular pose estimation SDK (e.g., Microsoft Kinect) is only effective in limited working conditions, so assuming depth cameras will always provide us with 3D poses is not valid.

Here, we present a weakly supervised solution to re-purpose existing state-of-the-art (SOTA) models for 3D pose estimation in novel practical applications when the use of MoCap systems for 3D data collection is not feasible. We make use of depth data as a higher-level and less precise estimation of 3D information to create heuristic rules and impose extra constraints on our MoCap-independent training data. Our final product is a predictive model which can estimate 3D human pose and its shape directly from a single 2D image. Our heuristic weakly-supervised 3D human pose estimation model (HW-HuP) is trained using both the learned pose and shape priors from a few public 3D pose datasets as well as 2D pose and depth observations from the target domain. By partially learning the priors from the source domain and the noisy observations from the target domain, our approach iteratively converge to a reliable 3D pose estimation. Depth data is employed during the supervised training, yet it is not required during the pose inference. This makes our HW-HuP approach different from the RGBD-based pose estimation algorithms which require full access to both RGB and depth modalities during the inference [9]. Our work addresses the problem of 3D human pose estimation for applications where no 3D pose ground truth can be collected in their natural settings by making the following contributions:

- Presenting a heuristic weakly supervised 3D human pose estimation (HW-HuP) approach that combines the priors from public 3D pose datasets and easy-to-access observations from the target domain in an iterative way.
- Targeting a practical application in the healthcare domain, in which an automatic 3D pose estimation of in-bed patients can lead to impactful outcomes. Since neither annotated 3D pose data no known prior are available, existing SOTA models performed poorly in this application as shown in Fig. 1, and there is an urgent need for a more reliable 3D pose approach such as HW-HuP.
- HW-HuP has flexibility in its input modality, such that when RGB is not informative (heavily occluded scenarios), other input modalities such as pressure map, depth, or infrared signals can be used in the pipeline.
- HW-HuP is evaluated on a large-scale public 3D pose dataset (Human3.6M [10]), quantitatively and on an in-bed pose dataset (SLP [6]), qualitatively.
- Among real-world applications of HW-HuP is infant motor development monitoring, where our approach leads to reliable infant 3D pose estimation, never solved before in the computer vision field, mainly due to its 3D data scarcity.

2 RELATED WORK

The 2D human pose estimation problem has been extensively studied in the past couple of decades based on 2D images, however when it comes to 3D pose estimation, addition 3D information is often necessary for model supervision, which usually comes from MoCap or depth sensing devices.

3D Human Pose Estimation using MoCap Data: 3D human pose estimation problem has been extensively studied in recent years. The 3D pose ground truth is mainly collected by motion capture system (MoCap) [10]. Early study conducted the pose estimation via a straightforward end-to-end learning, from an image to infer the 3D pose directly [11]–[15], as well as decoupling the 3D pose problem into 2D pose estimation problem with the 3D pose regression [16]–[19]. For example, in [20], a slim baseline model is proposed as 2D-to-3D lifting to directly regress the 3D pose from the given 2D landmarks.

Due to the limitation in variations in the public 3D human pose datasets, training on the existing 3D data alone is usually not enough to form a robust model. Accordingly, one way for estimation improvement is introducing additional supervision either by additional data, or additional constraints such as kinematic, skeletal, or temporal constraints [21]–[24]. In [21], 2D data is introduced into the model training combined with the symmetry and skeleton regulation. Though no additional 3D pose data is introduced, yet the large quantity of the available 2D human pose data provides sufficient variations to generalize the model for images in the wild. In [25], existing 3D pose data is further augmented via crossover and mutation to improve the model performance.

As more effective constraints are introduced, no 2D-to-3D correspondence is even needed during supervision which leads to the weakly supervised approaches. The key idea comes from learning the 3D pose priors from unpaired 3D pose data. A 3D pose dictionary is introduced in [26], as n [27], an independent network is employed to learn the pose prior distribution. In [18], [28], [29], the pre-trained low-dimensional representations from 3D is introduced to form prior in order to find plausible 3D poses. Kinematic constraints, anatomical loss, and joint limit constraints are also introduced for learning in [30]. Adversarial learning based approaches [21], [31], [32] although does not model the prior explicitly, the discriminator is inherently capable to determine if the pose is from the true prior or not. 2D and 3D temporal information is also integrated to recover the 3D pose with the geometry prior [33].

Another pipeline is by directly fitting the learned template [34] to a given image via optimization [35]. This approach directly regresses the pose and shape parameters from the given image, while the 3D pose data is still highly needed as the prior constraints. Among these works, either the 3D pose data is directly employed for supervision or to form a pose prior, it plays a key role to infer a reasonable 3D pose.

3D Human Pose Estimation using RGBD Data: Another plausible assumption is that when depth or RGBD data exists, the 3D pose data can be easily acquired, as many existing works seem to be able to successfully capture
human pose, even detailed body mesh using RGBD data [36]–[38]. Commercialized products even comes with their own software development kit (SDK) to directly estimate the 3D pose from the input depth data such as Microsoft Kinect. However, specific conditions have to be met for these approaches to work. When the major focus is on full body mesh reconstruction, continuous depth data is often required. Furthermore, specific poses for initialization, clean segmented point cloud, and slow motion are usually required in these works [39]–[42]. These mean such tracking has to be done in a controlled environment from a cooperative human subject. As for Microsoft Kinect SDK, it is only effective within a specific range and human body has to be clearly away from any surrounding walls.

These requirements could be challenging to be satisfied in many real-world applications. The data collected from a practical application may only have non-continuous frames [6], [7], as well as depth noise from items in the environment. Single frame depth estimation approaches do exist, which usually relies on a convolutional network for inference and show higher tolerance to noise and occlusions. In [43], the difficult 3D pose estimation problem is mapped to a simpler per-pixel classification problem and the confidence score of the body joints 3D proposal is generated from the reprojected classification result. In [44], a viewpoint invariant representation has been introduced to embed local regions into a learned viewpoint invariant feature space for pose estimation from varying view points. Anchor joint concepts have also been introduced in [45], where the joint location is regressed via the relevant anchor locations. These work although depth-based during inference, yet the ground truth 3D pose data or the segmentation are required during training [43]–[45]. In a previous work, we demonstrated that when they are tested on a real-world application with no 3D pose ground truth for training such as in-bed pose estimation, these pre-trained models are not performing well.

Unsupervised 3D Human Pose Estimation: There are few approaches that use no 3D pose data during training [19], [46]. In [46], the learning process relies on neither 3D pose data nor previous learned prior, nor multi-view requirement but only on 2D datasets which is impressive and seems promising for problems when 3D data is not available. However, the unsupervised learning inherently requires more data than a supervised approach and large quantity of 2D pose data is necessary for this approach. [19] proposed an unsupervised representation learning, yet the final decoder network still required some 3D pose data for supervision. Furthermore, the multi-view and video of the same subjects are required which are not always available or provided in a real-world application. With a multi-view setting, even though the very recent works in [47]–[49] successfully estimate 3D poses without any explicit 3D labels, the 3D coordinates of a space point given its 2D location can be recovered in a straightforward way. However, a calibrated multi-view setting is not always available in many applications.

Although innovative approaches keep emerging in the field and are deemed as general purpose, we have to note that the extensively used datasets which define such “general” domains share high similarity by focusing on a series of fixed daily activities. For example, in the existing 3D pose datasets [10], [50], only a few selected daily activities are collected such as walking, sitting, or jogging [51]. On the 2D part, the datasets are mainly collected from the web [52], [53], where such daily activities are more commonly seen. This actually explains why many approaches without 3D supervision but only learned priors are still effective on an novel input from other datasets for general purpose. A sharp contrast is when these well-performed models are applied on a specific application with different types of human subjects (e.g. infants) or activities (e.g. sleeping), where their superiority no longer holds or even ends in a total failure. A typical example is reported in [6], for bed-bound patient monitoring, that the SOTA 3D human pose models [21], [45], [54] cannot give a correct 3D inference on a relatively simple in-bed pose despite their outstanding performance on well-known benchmarks with more complicated poses [10].

Different from the majority of the 3D pose estimation work that are conducted under the “general” context with poses commonly seen in daily activities, this paper focuses on underrepresented cases with less or no ground truth 3D pose labels, that are often ignored by the main computer vision pipelines. With no 3D pose data and no learned prior from the target domain and only scarce discontinuous data, human 3D pose estimation could be a challenging yet rewarding task in many practical applications, such as in-bed patient or infant monitoring.

3 3D HUMAN POSE ESTIMATION WITHOUT 3D GROUND TRUTH OR LEARNED PRIOR

This paper focuses on enabling 3D human pose estimation in practical scenarios, in which not only the SOTA 3D pose estimation models do not perform well, but also due to the context-related constraints, no MoCap data can be collected to fine-tune the existing models with ground truth 3D pose data. Throughout this study, we assume 2D human pose data are available since even in very data-restricted settings, some manually labeled 2D pose data can be collected and used for conventional fine-tuning of the 2D pose estimation models [55]. Therefore, in our approach the following data availability is assumed: (1) 2D image data I, (usually RGB, unless otherwise); (2) depth data D; (3) the ground truth 2D pose annotation x; and (4) a pre-trained 3D pose estimation model F trained on large-scale publicly available 3D human pose datasets. Considering that we are targeting applications of 3D pose estimation in natural settings, we also have the following assumption on data constraints: (1) depth data D may not be continuous but provided as individual frames; and (2) D could be noisy and the subject may not be clearly segmented.

Uncertainty in Depth Observation: Given the 2D joints x and the depth map D, one may assume 3D pose can simply be achieved by combining the 2D pose and its projection from depth map. However, depth map only provides a series of surface points instead of the true joint locations. We call these points depth-based 3D proxy, which may have some biases from their true corresponding joint locations. If there were equal bias across all body joints, the true 3D pose could still be achieved by making everything root centered. However, these biases are not equal or constant and they
are both pose- and shape-dependant. A large limb does not necessarily hold the same bias as a thin limb due to their shape differences. The bias may also vary from different view points caused by pose differences.

If we assume all the limbs have an oval cross-section, in a single body case (e.g. human head), the bias variation mainly comes from an uneven distance of the joint to its surrounding surfaces which is usually small as shown in Fig. 2 (a). In a double body case (e.g. crossed arms or legs), the bias is mainly caused by the occlusion when the occluded joint is mapped to the upper surface of the other body part. Such error is usually large across limbs as shown in Fig. 2 (b) and needs to be avoided. An example of a double body bias that can occur in measuring the location in Fig. 2 (c). Therefore, the 3D joint location estimated based on the depth data is only a proxy from the true 3D pose.

4 INTRODUCING HW-HuP

The data distribution gap between the public MoCap data and a practical pose estimation application often stems from the differences in types of poses that the subjects take in different domains. As a result, the pre-trained SOTA models do not show satisfactory performance on a novel application unless being fine-tuned. Nonetheless, having no ground truth 3D pose data prevents a straightforward fine-tuning.

**Problem Formulation:** Our objective is to obtain a 3D pose regression model $F$ to estimate the human’s 3D pose $\theta$, its shape $\beta$, and a weak-perspective camera model $C$, directly from a single 2D image $I$, as $[\Theta, C] = F(I)$, where $\Theta = (\theta, \beta)$ represents the human pose and shape parameters. Following [56], [57], $\theta$ and $\beta$ are based on a shape parametric human template called SMPL [34]. Pose $\theta \in \mathbb{R}^{3K+3}$ is a vector of axis-angle representation of the relative rotation of $K = 23$ body parts with respect to its parent in the kinematic tree plus the root global rotation. Shape $\beta \in \mathbb{R}^P$ is the first $P = 10$ PCA coefficients in the human template space. SMPL model is a differential function that outputs a triangulated mesh $M(\Theta) \in \mathbb{R}^{3 \times N}$ with $N = 6980$ vertices. Camera model $C = (T, s)$ includes a translation vector $T \in \mathbb{R}^2$, and the scale $s \in \mathbb{R}$. For a 3D keypoint $X(\Theta) \in \mathbb{R}^3$, which is a linear combination of the human mesh vertices from $M$ [35], [57], its 2D projection $x$ is given as $x = s\Phi(RX(\Theta))+T$, where $\Phi$ is an orthographic projection, where $R$ is the global rotation matrix $R \in \mathbb{R}^{3 \times 3}$.

![Fig. 2. Depth proxy point bias from the true joint location in the case of: (a) single body with uneven shape, (b) double bodies with occlusion, and (c) right hip of a human body.](image)

**4.1 HW-HuP Framework Components**

HW-HuP learns 3D pose information jointly from the pose priors in the source domain as well as observations from the target domain. The observation in target domain will be interpreted while learning as the 3D regression model $F$ converges. An overview of our HW-HuP framework is shown in Fig. 3 which includes:

- **Two Networks:** (1) a 3D pose regression network $F$ to estimate the human pose $\theta$, shape $\beta$, and camera parameters $C$ from an input 2D image $I$; and (2) a differentiable neural renderer [58], which outputs the predicted depth $D$ and mask $Ma$ from the predicted human mesh $M(\theta, \beta)$.

- **Two Pipelines:** (1) an Optimization pipeline $OPT$, based on the SMPLify model [35], to minimize the 2D pose $x$ estimation error under the source priors based on the input initialization $\Theta_i$; and (2) an observation $OBS$ pipeline, which keeps feeding the observed 2D pose $x$, depth $D$, human body mask $Ma$ and the joint depth-based proxy $X_{dp}$ for supervision.

- **Two-Stage Supervision:** Our coarse to fine training paradigm has two stages. In the stage I, (1) from the input 2D image $I$, 3D regression network $F$ gives initial estimates on the 3D pose $\Theta_i$ and camera parameters $\hat{C}$. (2) $\Theta_i$ serves as the initialization for the optimization pipeline $OPT$. $OPT$ optimizes from current estimation under the source prior to have $\Theta_{opt}$, as a guess of current observation. (3) The regression network $F$ is updated by minimizing the error from observations $x$ and $X_{dp}$. After each iteration, regression model $F$ is updated to have better estimation $\Theta^*$. (4) When the model converges at stage I (by setting specific epoch), we enter stage II. In the stage II, we replace the noisy $X_{dp}$ supervision with an aligned depth $D$ supervision for a more detailed refinement.

**4.2 HW-HuP Key Features**

Here, we will describe novel features of the HW-HuP framework designed for 3D human pose estimation.
Selective Pose Prior Transfer: We learn the source priors via an optimization process, which employs a pre-trained Gaussian mixture model (GMM) prior [35]. Mixture of regression and optimization in a loop were originally introduced in [57], where the optimization serves as a refinement mechanism to update the regression output in the same domain. In our HW-HuP design, the optimization pipeline OPT mainly plays the role to introduce the source priors into the regressor $F$ with the supervision loss:

$$L_{OPT} = L_{2D} + \lambda_0 L_0(\theta) + \lambda_\beta L_\beta(\beta) + \lambda_\alpha L_\alpha(\theta), \quad (1)$$

where $L_{2D}$ is 2D pose loss, $L_0$ is GMM pose prior learned from the source domain, $L_\beta$ is a quadratic penalty on the shape coefficients, and $L_\alpha$ is the unnatural joint rotation penalty of elbows and knees. We have followed the same definition for each penalty term as in [35]. The optimization model is initialized with a strong source prior regulator, which degenerates over time with an exponential decay as $\lambda_0 = \lambda_{0,0} f^e$, where $\lambda_{0,0}$ is the initial weight for the GMM prior, $f$ is the decay factor, and $e$ is the epoch number. With the source prior faded away over time, the major body parts will follow the new target observation, yet the learned prior from source is kept for small limbs.

Around the idea of learning from the source prior, there are two plausible arguments:

- One may assume that learning from the source pose prior is pointless as there will be a different pose distribution in the target domain. However, we argue the incompleteness of the target pose distribution caused by the limited observations can be mitigated by using source priors. For example, we only have limited major joint locations from the target observation to regulate major body parts, and their axis rotation as well as small body parts are unconstrained. Since, the specific joint pose (rotation) priors are shared between the source and target, we can learn these information from the source priors (imagine that people do not rotate their head to extreme angles in any contexts).

- If we conduct transfer learning by using a pre-trained model, such prior is inherited from the learned weights. This is true if this transfer learning conducted within the same modalities. However, when we try to extend this model beyond RGB and without a good initialization, these prior via OPT will play an important role.

Taking the in-bed human pose for example [6], the underlying logic comes from the follow observation: if we look at the estimation result of the target image with a pre-trained model from a source domain [57], the output human usually holds a standing-like posture. Most of the times, we only need to change a few major joints to have an acceptable solution for the target observation. In other words, the estimated pose from the source model and the true pose in the target domain are neighbors on the human pose manifold with similarity in many trivial joint poses. A pose guess from a source prior to regress to the target pose is always better than a guess from nowhere and we hope the evolution follows such manifold. With a strong prior at beginning, the regressor $F$ learns under the source prior guidance to have a whole picture of full joint pose (rotation). With source prior faded away, $F$ will focus more on the observations to have its context-specific prior over the major joints yet keep the learned prior over the trivial ones which is otherwise unconstrained under the target-only supervisions. In this way, the source pose prior is actually selectively transferred to the target regressor $F$.

Joint Depth-based Proxy Supervision: As introduced in Section 5.2, the joint depth proxy $X_{dp}$ is noisy with bias depending on the human pose and shape and the bias could be significant or negligible depending on its causal source (as shown in Fig. 2). Since the significant biases mainly come from the occlusion cases, we explicitly employ the visibility information to filter out misleading invisible joints. Our weakly supervised 3D pose loss via $X_{dp}$ is given as:

$$L_{3D} = \sum_{j=1}^{K} V_j ||(X_{dpj} - \hat{X}_j)||^2_2, \quad (2)$$

where $V_j$ stands for the $j$-th joint visibility. Visibility are often provided in the 2D annotations [6], [52], [53], however the information is not usually used since most 2D pose estimation models attempt to infer all joints in the image no matter they are visible or not. Nonetheless, in $X_{dp}$ supervision, it is critical to avoid the significant bias in the observations. When the visibility information is missing (such as in an unlabeled dataset), we train an visible joint detection model by fine-tuning one of the SOTA pose estimation models (e.g. [59]) and adding an visible joint detection branch called VisNet, which is elaborated in detail in Section 5.2, on top of its backbone network.

Observation Interpretation while Learning HW-HuP supervision process is designed in a coarse to fine manner in two stages. At stage I, we supervise $F$ with only 2D pose $x$ data, the joint depth proxy $X_{dp}$, and the $\Theta_{opt}$ under source priors. In this stage, the trained model can yield a plausible front view of fitted model with aligned 2D poses. Combined with the $L_{3D}$ supervision with $X_{dp}$ after filtering, the estimated joints will also be close to their true locations $X$ in the 3D space. At stage II, despite the depth error, the estimated mesh from stage I has been already well aligned in 2D and the estimated mask $Ma$ can be further employed for depth alignment.

In our design, we employ a differential neural renderer (NR) [58] to generate the estimated depth $\hat{D}$ and human mask $Ma$ from estimated $\Theta_i$. We employ a straightforward strategy to directly match the depth in segmented area from the union of $Ma$ and the known mask of the data $Ma$. Getting exact $Ma$ requires additional fine-tuning and high quality segmentation annotation of boundary, which is time consuming. So, here we employ a weak mask to filter out the abnormal reading of $D$ of rough depth range and bounding box annotation. For the $z$ direction, we first estimate the root distance along $z$ from the rendered human mesh to the depth observation by minimizing the average distance from the observed depth to rendered depth in segmented area:

$$b_0 = \arg \min_b |D - \hat{D}| \cap (Ma \cap Ma), \quad (3)$$

where, $D$ and $\hat{D}$ stand for the ground truth depth and rendered depth, and $Ma$ for the estimated human body
The depth loss for stage II is given as:

\[ L_D = \| (D - b_0 - \hat{D}) \odot (Ma \cap Ma) \|_2^2. \]  \hfill (4)

Here, we employ the L2 norm for the depth as our depth is filtered and smoothed. In case of depth with strong noise or outliers, we recommend the robust penalty loss such as German-McClure [60]. In stage I, the joint is only pushed to their corresponding surface point via \( X_{dp} \). In stage II, by matching the correct surface point, the pose is further pushed into their true location. So, in stage II, the major mask of ground truth 3D pose annotation, in this experiment, we suppose to be able to infer occluded joints for a robust estimation work since (1) an effective pose estimation model irrelevant to its 2D visibility. However, as we discussed in Section 3, the visibility could be a potential indicator of large bias for joint depth proxy. Therefore, it is helpful to filter out unreliable proxy points during supervision. In our design, the VisNet head is based on the a ResNet [66] backbone. It includes two convolution layer with kernel \( 1 \times 1 \) and channel 256 and 32 followed by 3 fully connected layer with channel of 256, 64, 17. Each layer is followed by a batch normalization and a rectified linear unit (ReLU). To enhance its semantic understanding of the specific joint for visibility detection, we add the pose head [67] on top of the backbone for joint training. VisNet design is shown in Fig. 4(a). In our implementation, VisNet is trained with an Adam optimizer with learning rate 0.001.

VisNet for Invisible Joint Detection: The main purpose of the VisNet is to detect the invisible joints in an image. This information is often deemed as trivial in many pose estimation work since (1) an effective pose estimation model suppose to be able to infer occluded joints for a robust performance, and (2) in the 3D pose cases, the poses are collected from a MoCap device independently which is irrelevant to its 2D visibility. However, as we discussed in Section 3, the visibility could be a potential indicator of large bias for joint depth proxy. Therefore, it is helpful to filter out unreliable proxy points during supervision. In our design, the VisNet head is based on a ResNet [66] backbone. It includes two convolution layer with kernel \( 1 \times 1 \) and channel 256 and 32 followed by 3 fully connected layer with channel of 256, 64, 17. Each layer is followed by a batch normalization and a rectified linear unit (ReLU). To enhance its semantic understanding of the specific joint for visibility detection, we add the pose head [67] on top of the backbone for joint training. VisNet design is shown in Fig. 4(a). In our implementation, VisNet is trained with an Adam optimizer with learning rate of 0.001 with a total epoch of 80. Our implementation is trained on COCO dataset, which shows an area under curve (AUC) score of 93.3% on COCO validation split. The ROC curve is shown in Fig. 4(b).

Qualitative result of VisNet on COCO and Human3.6M [10] are shown in Fig. 5. We notice that it is more likely to find the occluded joint when the occluded object has distinct
5.3 Ablation Study

We choose SLP in-bed pose dataset to show a clear example, in which no prior pose data is ever known and no 3D pose data is provided. For a fair comparison with the other 3D pose estimation approaches [56], [57], we start with the RGB images, where human subjects are not heavily occluded, similar to the setting in the majority of the SOTA. We report aligned depth error for quantitative reference and also the qualitative comparison, which intuitively shows performance differences between varying configurations. For SOTA approaches, we choose the pre-trained model SPIN* [57], fine-tuned models, SPIN and human mesh recovery (HMR) model [56]. For our ablation, we form and compare several intuitive solutions for transfer learning in following settings:

- **3D-dp**: “Assuming 2D pose and depth data together are the same as the 3D pose, so just fine-tune the network with the $X_{dp}$ as the ground truth 3D pose.” For this setting, we train the network with only $x$ for 2D and $X_{dp}$ for 3D pose.

- **3D-dp-vis**: Following the same setting of 3D-dp, but adding the visibility constraint to rule out joint depth proxy with large biases.

- **3D-dp-vis-D**: HW-HuP full setting with depth alignment for stage II supervision.

- **noPrior**: “Assuming source prior is useless and misleading, so we only need to fine-tune the network with the target data.” For this setting, we set the source prior constraint to zero.

### Table 1

| SOTA          | HW-HuP Ablation |
|---------------|-----------------|
| SPIN* [57]    | 80.10           |
| SPIN [57]     | 68.38           |
| HMR [56]      | 63.43           |
| 3D-dp         | 48.13           |
| 3D-dp-vis     | 47.41           |
| 3D-dp-vis-D   | 36.01           |
| noPrior       | 38.54           |
| 2D-D          | 39.32           |

### 5.4 Extending Inputs Beyond RGB

Besides the commonly-used RGB modality, we further extend our model to work with the other imaging modalities and in more challenging conditions such as heavy occlusion.
and total darkness [6], [68]. In this experiment, we train HW-HuP on non-RGB modalities including depth, LWIR, and pressure map (PM) and their combination under all cover conditions provided in SLP noted as “nocover”, thin sheet as “cover1” and thick blanket as “cover2” [6]. Their performance in each cover condition and their overall performance are reported in Table 2. These results reveal that HW-HuP is still effective under these challenge conditions. Nonetheless, it is reasonable that the performance for “nocover” cases are noticeably better than the covered conditions. Please notice that aligned depth error is only a quantitative reference based on depth data alignment, which does not equal to their true 3D pose estimation performance. In this case, it is reasonable that the depth modality, who also serves as the metric shows better performance.

To provide a more comprehensive evaluation of their performance, we also provide the 2D evaluation result in PCK@0.2 metric [69], which means the threshold of acceptence is set to 20% of the human’s torso size. The PCK@0.2 results are shown in Table 3. Combined with this metric, we can see that the multimodal solution is still helpful to localize the joint location which agrees with the results in [6]. The outputs of HW-HuP on one of the occluded human images (from “cover2” condition) are shown in Fig. 7, which show that the results fit the observed point clouds nicely. More instances are exhibited in Appendix A.

### 5.5 Comparison with SOTA on a Large-Scale 3D Human Pose Benchmark

Due to the lack of 3D pose ground truth in the SLP dataset, to quantify HW-HuP performance, we directly apply HW-HuP on Human3.6M and compare its performance with the reported SOTA approaches, as shown in Table 4. From Table 4, we can see that HW-HuP differs from the oracle performer SPIN [57] by about 40mm before PA alignment, yet only 10mm after PA. Some predicted results are also exhibited in Fig. 8. It is apparent that the produced models not only have good 2D projection on the RGB images, they also match the point clouds effectively by only having access to the depth auxiliary information, without any 3D keypoints ground truth supervision (as shown in the last row of Fig. 8). As our main goal is fine-tuning the existing models for novel applications, HW-HuP is always initialized from the pre-trained SPIN [57], which benefits from daily activity tasks.

| Nocover | Depth | LWIR | PM | Combined | Cover1 | Depth | LWIR | PM | Combined |
|---------|-------|------|----|----------|-------|-------|------|----|----------|
| SPIN* [57] | 108.78 | 89.41 | 102.27 | 96.13 | SPIN* [57] | 105.82 | 92.65 | 101.01 | 102.27 |
| HMR [56] | 68.70 | 69.20 | 75.70 | 72.55 | HMR [56] | 72.81 | 71.25 | 75.92 | 72.52 |
| SPIN [57] | 62.72 | 66.60 | 70.28 | 67.35 | SPIN [57] | 67.71 | 67.81 | 70.29 | 67.96 |
| HW-HuP | 33.87 | 38.45 | 41.48 | 37.62 | HW-HuP | 36.87 | 39.82 | 41.43 | 37.92 |

| Cover2 | Depth | LWIR | PM | Combined | All Covers | Depth | LWIR | PM | Combined |
|--------|-------|------|----|----------|------------|-------|------|----|----------|
| SPIN* [57] | 105.34 | 89.31 | 101.08 | 100.82 | SPIN* [57] | 106.65 | 90.46 | 101.46 | 99.74 |
| HMR [56] | 73.06 | 70.69 | 76.10 | 73.11 | HMR [56] | 71.52 | 70.38 | 75.91 | 72.73 |
| SPIN [57] | 68.38 | 66.58 | 70.40 | 68.15 | SPIN [57] | 66.27 | 67.00 | 70.32 | 67.82 |
| HW-HuP | 37.26 | 40.42 | 41.68 | 38.26 | HW-HuP | 36.00 | 39.56 | 41.53 | 37.93 |
TABLE 3
PCK@0.2 of HW-HuP and SOTA approaches for 2D ground truth and 2D results projected by 3D predicted pose on the “nocover”, “cover1”, “cover2” and all covers conditions of the depth, LWIR, PM from SLP dataset.

|                | Nocover         |                  |                  |                  | Cover1        |                  |                  |                  |
|----------------|-----------------|-----------------|-----------------|-----------------|----------------|-----------------|-----------------|-----------------|
|                | Depth | LWIR | PM | Combined | Depth | LWIR | PM | Combined | Depth | LWIR | PM | Combined |
|                | SPIN* [57]  | 34.43 | 42.85 | 15.49 | 45.90 | SPIN* [57]  | 19.63 | 17.21 | 15.82 | 24.58 |
|                | HMR  [56]  | 96.51 | 95.57 | 90.59 | 95.46 | HMR  [56]  | 90.76 | 91.93 | 90.71 | 94.14 |
|                | SPIN  [57]  | 96.23 | 95.57 | 90.56 | 95.45 | SPIN  [57]  | 90.37 | 92.21 | 90.51 | 94.17 |
|                | HW-HuP | 96.64 | 95.22 | 91.19 | 95.49 | HW-HuP | 91.70 | 91.30 | 91.20 | 94.21 |

5.6 Qualitative Analysis on Another Real-World Application: Infant 3D Pose Estimation

Infant motion analysis is a topic with critical importance in early childhood developmental studies [77]–[79]. The infant’s motor activities in her natural home environment can be captured conveniently through a baby monitor and process based on the infant’s change of pose over time. However, due to the lack of publicly available infant pose data (caused by privacy and security considerations), there are very few work tackling infant pose estimation in the computer vision community. Authors in [80] collect an infant dataset dubbed as MINI-RGBD, however the focus of their work is more on infant mesh generation and the pose diversity in MINI-RGBD is limited to a very few infants during their scan sessions. To the best of our knowledge, there is no existing trained model, neither a known prior for infant 3D pose estimation, and given its novel context, it makes an ideal use-case for our HW-HuP approach. To this end, we use two infant pose datasets to train and validate HW-HuP performance: (1) The public synthetic and real infant pose (SyRIP) dataset [81], which combines the real (scraped from web) and synthetic infant data for our 2D supervision. We keep the last 100 real images for test purpose and 1600 for the training session. (2) The private modeling infant motor movement (MIMM) dataset, which is collected by our collaborators (as part of an NIH funded study) and contains video recordings of motor assessment sessions from 68 infants under one year of age in an interactive setting with their caregivers and clinical assessors. In MIMM dataset, the depth data from an MS Kinect camera is provided, yet no 3D pose is available. We keep the last two infant data for test purpose. We initialize HW-HuP with the pre-trained model of [57] and train on the these hybrid infant data for 640 epochs. The aligned depth error on MIMM is 42.0mm. The qualitative results for test samples from both MIMM and SyRIP are shown in Fig. 9, which shows that HW-HuP can capture the infant 3D pose in both clinical environment and in the wild.

6 Conclusion

In this work, we propose a transfer learning strategy for 3D human pose estimation when neither ground truth 3D annotations nor learned prior are available in a novel application with limited data. By selectively learning from the source priors as well as a series of easy-to-access target observations, our HW-HuP model yields a robust 3D pose estimation performance even under challenging target contexts. As a practical example application, we employed HW-HuP on an in-bed pose estimation problem, where other SOTA approaches failed. We also demonstrated HW-HuP flexibility with other non-RGB imaging modalities for solving 3D human pose estimation problem in heavily...
TABLE 4
Comparison with the SOTA based on the MPJPE metric tested on Human3.6M dataset using Protocol#2. We also report the result of rigid Procrustes analysis (PA) alignment here [62] for reference. We can see that without PA, our performance is already comparable to some of the SOTA approaches despite not using any ground truth 3D pose in our training pipeline. SPIN can be deemed as an oracle for us, which employs the same network structure as HW-HuP, yet with access to 3D pose ground truth. Best MPJPE is given in bold and PA MPJPE in italic.

| Methods | Dir. | Dis. | Eat | Gre. | Phon. | Pose | Pur. | Sit | StD. | Smo. | Phot. | Wait | Walk | WalkD. | WalkT. | Avg |
|---------|------|------|-----|------|-------|------|------|-----|------|------|-------|------|------|--------|--------|-----|
| Akhter & Black [70] | 199.2 | 177.6 | 161.8 | 197.8 | 176.2 | 186.5 | 195.4 | 167.3 | 160.7 | 173.7 | 177.8 | 181.9 | 176.2 | 198.6 | 192.7 | 181.1 |
| Ramakrishna [71] | 137.4 | 134.9 | 141.6 | 154.3 | 157.7 | 158.9 | 141.8 | 158.1 | 168.6 | 175.6 | 160.4 | 161.7 | 150.0 | 174.8 | 150.2 | 157.3 |
| Zhou [72] | 99.7 | 95.8 | 87.9 | 116.8 | 108.3 | 107.3 | 93.5 | 95.3 | 109.1 | 137.5 | 106.0 | 102.2 | 106.5 | 110.4 | 115.2 | 106.7 |
| SMPLify [35] | 62.0 | 60.2 | 67.8 | 76.5 | 92.1 | 77.0 | 73.0 | 75.3 | 103.3 | 137.3 | 83.4 | 77.3 | 79.7 | 86.8 | 81.7 | 82.3 |
| Chen [17] | 89.9 | 97.6 | 90.0 | 1079 | 107.3 | 93.6 | 136.1 | 133.1 | 240.1 | 106.7 | 139.2 | 106.2 | 87.0 | 114.1 | 90.6 | 114.2 |
| Tome [28] | 65.0 | 73.5 | 76.8 | 86.4 | 86.3 | 68.9 | 74.8 | 110.2 | 173.9 | 85.0 | 110.7 | 85.8 | 71.4 | 86.3 | 73.1 | 84.8 |
| Moreno [73] | 69.5 | 80.2 | 78.2 | 87.0 | 100.8 | 76.0 | 69.7 | 104.7 | 113.9 | 87.9 | 107.2 | 98.5 | 79.2 | 82.4 | 77.2 | 87.3 |
| Zhou [33] | 68.7 | 74.8 | 67.8 | 76.4 | 76.3 | 84.0 | 70.2 | 88.0 | 113.8 | 78.0 | 98.4 | 90.1 | 62.6 | 75.1 | 73.6 | 79.9 |
| Jahangiri [74] | 74.4 | 66.7 | 67.9 | 75.2 | 77.3 | 70.6 | 64.5 | 95.8 | 123.7 | 79.6 | 79.1 | 73.4 | 67.4 | 71.8 | 77.6 | 77.6 |
| Mehta [11] | 57.5 | 68.6 | 59.6 | 67.3 | 78.1 | 56.9 | 69.1 | 98.0 | 117.5 | 69.5 | 82.4 | 60.0 | 53.3 | 63.5 | 59.1 | 61.4 |
| Martinez [20] | 51.8 | 56.2 | 58.1 | 59.0 | 69.5 | 55.2 | 58.1 | 74.0 | 94.6 | 62.3 | 78.4 | 59.1 | 49.5 | 65.1 | 52.4 | 62.9 |
| Fang [75] | 50.1 | 54.3 | 57.0 | 57.1 | 66.6 | 53.4 | 55.7 | 72.8 | 88.6 | 60.3 | 73.3 | 57.7 | 47.5 | 62.7 | 50.6 | 60.4 |
| Sun [76] | 52.8 | 54.8 | 54.2 | 54.3 | 61.8 | 53.1 | 53.6 | 71.7 | 86.7 | 61.3 | 67.2 | 53.4 | 47.1 | 61.6 | 63.4 | 59.1 |
| Sun [75] | 47.5 | 47.7 | 49.5 | 50.2 | 51.4 | 43.8 | 46.4 | 58.9 | 65.7 | 49.4 | 55.8 | 47.8 | 38.9 | 49.0 | 43.8 | 49.6 |
| Moon [54] | 50.5 | 55.7 | 50.1 | 51.7 | 53.9 | 46.8 | 50.0 | 61.9 | 68.0 | 52.5 | 55.9 | 49.9 | 41.8 | 56.1 | 46.9 | 53.3 |
| SPIN [57] | 60.6 | 61.0 | 57.9 | 64.5 | 67.1 | 59.9 | 57.7 | 80.1 | 91.3 | 63.2 | 63.8 | 60.2 | 33.2 | 61.2 | 60.4 | 65.7 |
| SPIN PA | 39.5 | 42.1 | 41.1 | 44.8 | 46.6 | 38.6 | 38.8 | 60.8 | 68.5 | 45.2 | 46.2 | 41.4 | 36.0 | 45.0 | 40.9 | 44.2 |
| HW-HuP | 95.0 | 103.9 | 99.8 | 101.0 | 106.3 | 94.2 | 110.1 | 121.4 | 140.1 | 103.3 | 112.8 | 99.5 | 92.8 | 107.2 | 101.9 | 104.1 |
| HW-HuP PA | 47.0 | 49.5 | 49.2 | 51.2 | 52.1 | 45.0 | 48.4 | 64.7 | 75.4 | 49.9 | 54.3 | 47.8 | 42.6 | 53.3 | 48.8 | 50.4 |

Fig. 9. Qualitative 3D human pose and shape estimation results of our HW-HuP applied on two infant datasets: MIMM and SyRIP. The first column shows the input RGB image. The second and third columns visualize the HW-HuP results of front view and side view, respectively. Point clouds are aligned with side view model in blue dots for MIMM result. Point clouds are not provided in SyRIP dataset and is not shown in side view.

ocluded scenarios and under total darkness. Moreover, we reported HW-HuP performance on a popular 3D human pose benchmark, Human3.6M, which showed comparable results with the SOTA approaches that have full access to the 3D ground truth pose data. By simply depending on an easy-to-deploy off-the-shelf depth camera, HW-HuP is able to solve many practical 3D human pose estimation problems, where the use of MoCap is infeasible under the application natural settings. HW-HuP enables patient monitoring in medical facilities as well as real-life human behavior studies in highly constrained spaces, such as infant movement in her bassinet or a crib, pilot training in the cockpit or driver behavior or gesture recognition inside a car.

7 ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 1755695. We would like to also thanks our collaborator at Early Markers to give us access to the modeling infant motor movement (MIMM) dataset as a testbed to showcase our algorithm on.

REFERENCES

[1] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, “Boosting for transfer learning,” in Proceedings of the 24th international conference on Machine learning. ACM, 2007, pp. 193–200.
[2] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, “Large-scale video classification with convolutional neural networks,” in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2014, pp. 1725–1732.
[3] O. Köpükli, M. Babae, S. Hörmann, and G. Rigoll, “Convolutional neural networks with layer reuse,” in 2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019, pp. 345–349.
[4] D. S. Maitra, D. Papava, V. Olaru, and C. Sminchisescu, “Human3.6m: Large scale datasets and predictive methods for 3d human pose estimation and machine intelligence,” IEEE transactions on pattern analysis and machine intelligence, vol. 36, no. 7, pp. 1325–1339, 2013.
APPENDIX A
MORE VISUALIZATION RESULTS

Here, we exhibit more 3D prediction examples of both SLP dataset in Fig. S1 and Human3.6m dataset in Fig. S2. In Fig. S1, we display predicted results based on different inputs (depth, LWIR, PM, or multi-modal) for subjects with different cover conditions (no cover, a thin layer sheet, or a thick blanket). Some failure cases are included such as row(r) 1, column(c) 1 where LWIR failed and also the r3c1 where the left hand is not in a rest pose. For visualization of Human3.6m frames, we observe that although the performance of our HW-HuP model is comparable to the SOTA 3D pose estimation models, there are still some failure cases. For instance, the predicted upper limbs are more difficult to align to the point cloud than the lower limbs (e.g. the 2nd pose of subject 9 and the 4th pose of subject 11 in Fig. S2).
Fig. S1. Qualitative 3D human pose and shape estimation results of our HW-HuP applied on SLP dataset when input 2D images are depth, LWIR, PM, or their combinations, respectively. First row shows the input modalities as well as a "nocover" version of the RGB image as the reference. Second and third rows show the inference result of front view and side view, respectively.
Fig. S2. More qualitative 3D human pose and shape estimation results of our HW-HuP applied on Human3.6m valid dataset. First three rows show the predictions of frames for subject 9. Last three rows are for the subject 11.