In the Wild Human Pose Estimation Using
Explicit 2D Features and Intermediate 3D Representations

Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Gerard Pons-Moll, Christian Theobalt
Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrucken, Germany
{ihabibie, wxu, dmehta, gpons, theobalt}@mpi-inf.mpg.org

Abstract

Convolutional Neural Network based approaches for monocular 3D human pose estimation usually require a large amount of training images with 3D pose annotations. While it is feasible to provide 2D joint annotations for large corpora of in-the-wild images with humans, providing accurate 3D annotations to such in-the-wild corpora is hardly feasible in practice. Most existing 3D labelled data sets are either synthetically created or feature in-studio images. 3D pose estimation algorithms trained on such data often have limited ability to generalize to real world scene diversity. We therefore propose a new deep learning based method for monocular 3D human pose estimation that shows high accuracy and generalizes better to in-the-wild scenes. It has a network architecture that comprises a new disentangled hidden space encoding of explicit 2D and 3D features, and uses supervision by a new learned projection model from predicted 3D pose. Our algorithm can be jointly trained on image data with 3D labels and image data with only 2D labels. It achieves state-of-the-art accuracy on challenging in the wild data.

1. Introduction

Human motion capture has a wide range of applications in computer animation and also other areas such as biomechanics, medicine, and human-computer interaction. However, the standard 3D human motion capture systems typically require marker suits and/or multiple cameras recording in a controlled setting which are expensive and complicated to set up, and are impractical outside of the lab or studio environments. Methods that infer 3D pose only from monocular images overcome many such limitations and make 3D pose estimation more widely applicable. However, due to the under-constrained nature of monocular 3D pose estimation, achieving accurate 3D prediction is still a challenging task.

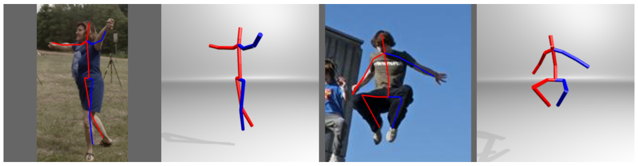

Figure 1: 3D pose prediction using our method for general scenes. Please refer to Sec. 3 for the details of our method and Sec. 4 for results and evaluations.

Recent progress of Convolutional Neural Networks (CNN) [17] has enabled promising learning-based methods for 3D human pose estimation from a single color image. Training such methods typically requires a large amount of RGB images annotated with reference 3D poses from either marker-based or markerless multi-camera motion capture systems [31, 8, 28, 15], synthetic data [6], or IMU-based systems [10, 38, 37]. Owing to this complex reference data capturing, diversity in real world appearance or pose is hard to achieve in training data, which limits the generalization of trained networks on in-the-wild scenes.

To improve in-the-wild generalization, previous work has leveraged features learned on in-the-wild annotated 2D pose data. Some methods [21, 22] proposed to finetune this learned representation on 3D pose prediction using 3D pose datasets captured in a studio. Others [43] use this learned representation as an initialization to jointly predict both 2D key points and depth information. For images where the 3D annotations are available, both 2D keypoints and depth are supervised, with supervision coming from geometric constraints otherwise. In this way, networks carry over features useful for in-the-wild 2D for better 3D pose estimation in out-of-studio settings.

Using a strong pre-existing pose prior, like a parametric body model, can also help a network to predict more accurate 3D poses if labelled 3D training data is scarce [41, 16].

Since 3D pose labels on general scene images are hard to obtain while larger annotated 2D training corpora exist, several deep learning based methods resort to using 2D pose as the target prediction, followed by an additional 3D pose lifting step [11, 39, 42, 2, 34, 5, 20]. Using such, [20] showed that 2D pose data alone is enough to train a network that achieves promising 3D pose estimation accuracy. However, solely predicting 3D from 2D pose is an inherently ambiguous task and in these approaches important 3D pose cues from the image are neglected.

In this paper, we introduce a new convolutional neural network architecture for 3D pose estimation that achieves state-of-the-art accuracy on challenging in-the-wild data. It introduces two main innovations that enable us to effectively train the network using both, more scarcely available image data with 3D annotation and more easy to generate image data with only 2D annotation.

The first innovation is inspired by 2D-to-3D pose lifting [20], but maintains the network's capability to explicitly utilize 3D cues in images. To this end, we design some channels of the convolutional latent space to encode explicit 2D keypoint features in heatmaps, leaving the rest of the features to contain "depth" information about the human pose. Separating the 2D and depth, and supervising 2D with additional in-the-wild data, which has been the primary driver of accurate 2D pose estimation methods [40, 23, 4], allows the network to consequently predict 3D pose more reliably even under a significant shift of the input appearance between the training and testing time. These 2D pose features can be trained jointly with depth features on data with 3D annotations, or trained independently on data with 2D annotations, while in both cases improving overall network performance.

The second innovation is a supervision approach that reduces 3D-to-2D ambiguity when training on data with 2D annotations only. To this end, we design a neural network that learns how to estimate the location of 2D body joints by using the 3D human pose predicted from the earlier network layers as latent features. More specifically, we learn to predict the weak perspective camera parameters of the given monocular image input that project the predicted 3D pose to the 2D space. During training, this projection loss can be used to update the information of 3D joint positions regardless if the training image has 3D labels or only 2D labels.

Our approach achieves a state-of-the-art accuracy of 70.4% 3D PCK on the MPI-INF-3DHP benchmark with challenging outdoor scenes, even when trained only using images with 3D pose labels from the H3.6M [12] studio dataset. When jointly training on larger corpora of in-studio images with 3D labels and in-the-wild data with 2D labels, we achieve 91.3% 3D PCK on MPI-INF-3DHP which outperforms all previous methods.

2. Related Work

Human pose estimation is an actively studied area in computer vision. We focus our discussion on recent learning-based approaches that are relevant to our work.

3D pose from 2D keypoint detection. Due to the robustness of some recent CNN-based 2D pose detection methods [36, 35, 40, 23, 4], many 3D pose estimation method reformulate the task as a combination of 2D keypoints prediction and body depth regression. Mehta *et al.* [22] combine 2D heatmap prediction with 3D location maps to estimate the position of each joint in the 3D space. Zhou *et al.* [43] propose a weak supervision training scheme using a stacked hourglass network [23] on both in-the-wild 2D data and studio data with 3D labels. The network is trained to predict 2D pose on both studio and outdoor dataset and at the same time also learns to predict depth information from the 3D labeled data. Yang *et al.* [41] also use similar weak supervision, but they extend this idea by introducing an adversarial network that learns how to differentiate between a ground truth and a predicted pose generated by the 3D pose prediction network. Another similar line of work is proposed by Dabral *et al.* [7] which improves this approach further by using body symmetry constraints and a separate temporal prediction network to achieve better 3D prediction stability across sequential frames.

To take the full advantage of the detection-based method, Pavlakos *et al.* [26] proposed using a volumetric representation as an extension of the 2D joint heatmaps in the 3D space. However, this formulation is computationally expensive to perform even after using their coarse-to-fine strategy proposed to mitigate this issue.

Direct 3D pose prediction. Instead of using the combination of 2D and depth prediction, several works regress 3D body keypoints directly. Tekin *et al.* [33] enhance a direct 3D prediction network by learning human body structure using a pose autoencoder.

Mehta *et al.* [21] use multiple intermediate supervision tasks, such as predicting the output at multiple network levels and predicting 2D heatmaps as an additional objective. They use two step training approach to improve generalization. The network is firstly trained to learn 2D joint heatmaps and then refined on the task of directly predicting 3D joint location maps from 3D annotated studio data. Instead of directly predicting the keypoints, [44] regresses the joint angles on a kinematic body model, assuming that the bonelength of the subject is known. Sun *et al.* [32] use a geometry aware formulation that also predicts bone length and bone vector orientation instead of only regressing 3D keypoint locations.

Rhodin *et al.* [30] proposed a multi-view consistent prediction approach during training to refine neural network's monocular pose prediction on general scenes. But it requires synchronized multi-camera footage to train. Multi-view settings can also be used to perform unsupervised or semisupervised learning on human pose estimation by training the network to learn a geometry aware latent space that

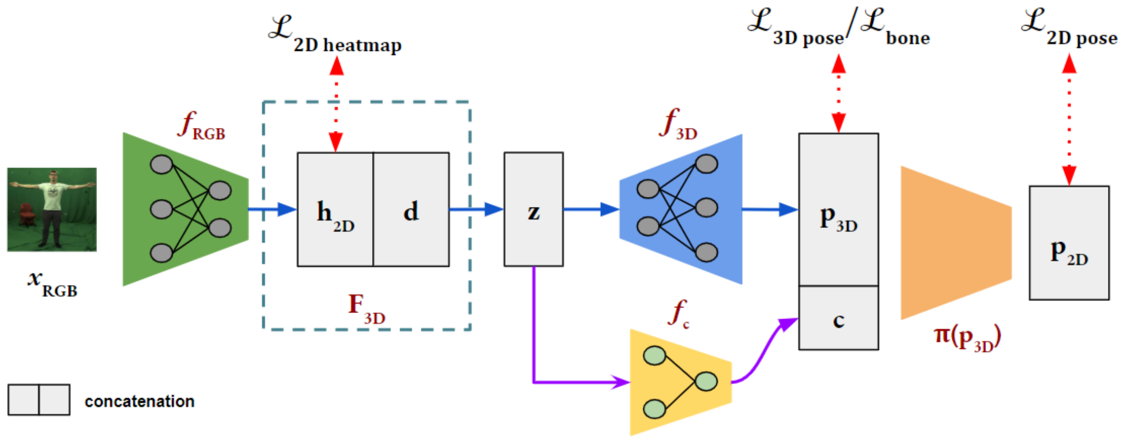

Figure 2: The overview of our proposed architecture. We use a CNN f_{RGB} to learn 3D pose features represented as 2D heatmap locations $\mathbf{h}_{2D}$ and additional 3D pose cues $\mathbf{d}$ in the latent space. Both information are used to predict a root centered 3D pose $\mathbf{p}_{3D}$ and viewpoint parameters $\mathbf{c}$ using networks f_{3D} and f_c, respectively. Finally, we concatenate $\mathbf{p}_{3D}$ and $\mathbf{c}$ to learn 2D keypoint information $\mathbf{h}_{2D}$, allowing the network to update 3D pose information even if 3D labels are not available.

can generate novel view on different cameras [29].

3D lifting without depth information. Some methods compute 3D pose by estimating the depth from the detected 2D keypoints only. Tome *et al*. [34] performs a sequence of 3D lifting and reprojection to iteratively improve prediction quality. Chen *et al*. [5] find a closest 3D pose from a library of human poses that best matches the detected 2D pose. [20] use a fully connected neural network with residual connection can achieve accurate 3D pose estimation performance using 2D ground truth or a very accurate 2D keypoint detection as input. Regardless, these approaches cannot overcome the principled ambiguity that there are many possible 3D body pose that can be correctly projected into the corresponding 2D pose. To reduce this ambiguity of 3D lifting from 2D estimates, Pavlakos *et al*. [25] use ordinal depth annotation between joint pairs, which is a special case of posebits introduced by Pons-Moll *et al*. [27].

Estimating 3D pose using 2D projection information. Bogo *et al*. [2] fit the 2D keypoints projection of the parametric SMPL [18] body model to 2D predictions from a separate method using an optimization approach. Brau *et al*. [3] demonstrated that 2D projection, body pose prior, and body part length information can be used as the training loss objectives for 3D pose prediction. Our method extends the idea of [3] by introducing additional 3D supervision and paired training on in-the-wild dataset. Kanazawa *et al*. [16] showed that pose and shape parameters of the SMPL body model from monocular images can be learned using a neural network. While their method uses a 2D projection loss of the body model as the main objective, their method also requires an adversarial regularizer against parametric body models. This method can be further improved by using additional labels of 3D pose and SMPL parameters if available. Omran *et al*. [24] proposed another deep learning approach to infer the parameters of the SMPL body model, and analyzed performance when varying the input representation (silhouettes, 2D keypoints, part segmentations) and the proportion of 2D and 3D data. Our approach outperforms these methods on several benchmark data sets.

The above review shows that many methods tackle generalizability on the in-the-wild images using either transfer learning from 2D pose task, or by decoupling the 3D pose estimation into separate 2D keypoint detection and depth regression problems. For methods that decouple the 3D representation [43, 41, 7], depth information is predicted if 3D labels are available and otherwise some weak supervision constraints (e.g. a parametric body model) are used for regularization. In this paper, we propose a new architecture that combines explicit encoding of separate 2D and 3D depth features in hidden space, instead of operating on vectorized 2D predictions as in previous lifting schemes. Our trained projection network further stabilizes overall 3D prediction accuracy.

3. Approach

The method estimates the root (pelvis) relative 3D locations of K human body joints $\mathbf{P} = \{\mathbf{J}_1, \ldots, \mathbf{J}_K\}$ in the camera reference frame from a monocular RGB image. Our method assumes that a crop around the subject is available.

A baseline strategy for our goal would be as follows: Given a training set consisting of pairs of RGB images and their corresponding 3D pose labels $\mathcal{D} = \{(\mathbf{I}_n, \mathbf{P}_n^{GT})\}_{n=1}^{N}$, we could train a convolution-based neural network $f_{RGB}(\mathbf{I}_n, \theta)$ to predict a vectorized representation of 3D joint locations. Network parameters θ could be trained by minimizing the difference $\mathcal{L}_{3D}$ between pose prediction and ground truth

$$\mathcal{L}_{3Dpose} = \frac{1}{N}\sum_{n=1}^{N} \parallel f_{RGB}(\mathbf{I}_n, \theta) - \mathbf{P}_n^{GT} \parallel_2^2 \qquad (1)$$

By training on currently available image data sets with 3D pose annotation, such direct supervision approach can already enable the network to achieve reasonable performance on studio test images. However, such a baseline method is still constrained in its ability to generalize to in-the-wild scenes due to the limited amount of available real world images with ground truth 3D poses.

We therefore introduce several strategies to augment such a 3D pose network such that it performs better on in-the-wild scenes. Our augmented network can be trained on both, images with 3D labels and in-the-wild images with only 2D labels. First, using an explicit 2D pose representation in the feature space of the CNN combined with 2D pre-training can significantly boost the quality of the prediction. Second, we propose additional supervision by using a trained projection sub-network that learns weak perspective camera information for projecting 3D pose estimates to the 2D image space. The overview of our network is shown in Figure 2.

3.1. Explicit 2D feature representation for 3D pose prediction

Martinez *et al*. [20] showed that a simple neural network is capable of directly regressing 3D human pose with good accuracy by using only vectorized 2D pose as input. This shows that a neural network is able to estimate the structure of natural 3D human pose from corresponding 2D information to some extent. However, such a lifting scheme can only remedy to some extent the fundamental ambiguity that multiple 3D poses can look the same in 2D. [25] showed that additional weak ordinal depth supervision can partially resolve the ambiguity of the problem.

We argue that a 2D-to-3D lifting approach can also be applied on 2D heatmap input instead of the vectorized 2D pose representation. From this observation, we decided to design the convolutional features of our CNN to explicitly encode 2D pose heatmap information. The idea behind this decision is to explicitly decouple 2D pose information from other learned features in the convolutional latent space. The rest of the feature maps can be used by the network to capture other image information related to 3D human pose, such as 3D depth. In this way, the network is guided to learn 3D pose features that are more reliable due to the robust 2D pose prediction and easier to interpret. Furthermore, by using a 2D training loss on this component we allow network to learn useful features from images when 3D pose labels are not available.

To this end, we design a convolutional feature map $\mathbf{F}_{3D} = [\mathbf{h}_{2D}, \mathbf{d}]$ after the extractor network f_{RGB}. This feature map consists of 64 output channels with a spatial dimension of 16×16. We use the first 14 channels to capture the 2D pose information. We optimize this region during training by minimizing the loss compared to the 2D ground truth heatmap in a least square sense. The rest of the feature channels $\mathbf{d}$ are not directly constrained by any explicit loss and will be supervised through the 3D pose, 2D projection, as well as additional pose constraints losses explained later.

To infer 3D pose from $\mathbf{F}_{3D}$, we first combine explicit 2D heatmaps $\mathbf{h}_{2D}$ and the additional features $\mathbf{d}$ learned by the convolutional encoder by using a simple fully connected layer into a latent vector $\mathbf{z} \in \mathbb{R}^{1024}$. Then, a fully connected network with residual connections f_{3D} is used to learn the vectorized 3D pose representation $\mathbf{p}_{3D}$. We design f_{3D} to be similar to the lifting architecture in [20]. More specifically, we use a series consisting of four fully connected layers with the width of 1024 and ReLU activations. A residual connection is also incorporated to connect $\mathbf{z}$ with the output of the second layer of f_{3D}.

Bone loss Several earlier works reported that detection-based approaches using a heatmap or volumetric representation tend to achieve better performance on both 2D and 3D pose estimation tasks than approaches regressing vectorized predictions. However, additional structure aware supervision can lift performance of vectorized prediction to a competitive level [32]. Since our method also performs vectorized 3D pose prediction, we complement the 3D training loss $\mathcal{L}_{3Dpose}$ (equation 1) with a bone supervision loss $\mathcal{L}_{bone}$. For 3D training data, $\mathcal{L}_{bone}$ measures the similarity of the vector between a joint $\mathbf{J}_k$ to its corresponding parent in the kinematic chain to ground truth. For 2D data, it measures the difference of scalar bone lengths to ground truth.

3.2. Predicting 2D projection from 3D pose

To further improve our method's ability to utilize 2D pose data for training 3D pose prediction, we train a sub-network to project the predicted 3D pose to the image space. Our camera network f_c predicts the principal coordinate (c_x, c_y) and the focal length (α_x, α_y) parameters of a weak perspective camera model from the given input image. By using the features extracted from the latent representation $\mathbf{z}$, we use a multi-layer perceptron to infer the camera parameters $\mathbf{c} \in \mathbb{R}^4$. During training, a 2D loss $\mathcal{L}_{2Dpose}$ measures the L2 distance between ground truth 2D pose and 2D projection $\mathbf{p}_{2D}$ of the predicted 3D pose:

$$\mathbf{p}_{2D} = \begin{bmatrix} \pi_x(\mathbf{p}_{3D}) \\ \pi_y(\mathbf{p}_{3D}) \end{bmatrix} = \begin{bmatrix} \alpha_x \mathbf{p}_{3D}(x) + c_x \\ \alpha_y \mathbf{p}_{3D}(y) + c_y \end{bmatrix} \qquad (2)$$

Our projection formulation allows the network to learn partial information about the 3D pose even when only 2D pose annotations are available. However, there are no constraints that guarantee the correctness of the predicted depth information. To regularize 3D pose prediction when training on 2D data, we use the additional bone loss $\mathcal{L}_{bone}$ enforcing bone length similarity to ground truth for additional supervision. We randomly pick the bone length of one of the training subjects as the ground truth for every training instance.

3.3. Network design

We use an adapted ResNet-50 [9] as the basis of the backbone subnetwork f_{RGB} (Figure 2) that extracts pose features from 2D images. This offers a good trade-off between prediction accuracy and inference time, allowing our network to be optionally used in real-time applications. The original ResNet-50 architecture is used up to level *Res4f* and we train level *Res5a* from scratch without striding while also reducing its number of output channels to 1024. This extractor network is then followed by the 3D pose regressor network described in 3.1.

The studio datasets with 3D labels and the outdoor data sets with 2D labels which we train on tend to have slightly differing image statistics due to contrast differences, as well as foreground background augmentations on the 3D data sets. To further mitigate this residual domain gap beyond what our new network architecture can already do by its design, we employ a similar pre-training approach as several earlier 3D pose prediction methods, e.g. [21]. To this end, we first pre-train our ResNet-50 network on ImageNet features to perform 2D heatmap prediction only. Here, intermediate 2D pose supervision is used on the first 14 channels of the *res4d* and *res5a* feature maps. The same intermediate supervision is also used later when finetuning the complete network on both 2D and 3D pose data. After pre-training, final training of the full network on both outdoor images with 2D annotations and studio images with 3D annotations results in learned features that generalize well to in-the-wild scenes and yield high accuracy in 3D pose estimation.

Our algorithm can be modified to handle input images of arbitrary framing around the human, because our subnetwork f_{RGB} is convolutional. For example, we can perform tight bounding box cropping around the detected 2D keypoints before passing the rescaled image into the subsequent sub-network.

4. Experiments and Discussion

After discussing datasets and network training we will show the high performance of our method qualitatively and quantitatively. We use the H3.6M data set [12] to compare general 3D pose estimation accuracy on in-studio data, and show that we outperform previous methods on the more general MPI-INF-3DHP benchmark set. The latter features more diverse motions, and more diverse scenes, including indoor scenes with green screen background (**GS**), as well as more in-the-wild scenes with general backgrounds, both indoors (**No GS**) and outdoors (**Outdoor**). An ablation analysis shows the significance of the individual components in the proposed approach. The supplementary document contains further explanations and evaluations.

Method	**PCK GS**	**PCK No GS**	**PCK Outdoor**	**PCK All**	**AUC All**	**MPJPE All**
Mehta [21]	84.6	72.4	69.7	76.5	-	-
Mehta [22]	-	-	-	76.6	40.4	124.7
Dabral [7]	-	-	-	76.7	39.1	103.8
Ours (US)	87.8	80.2	73.8	81.5	44.5	90.7
Ours (GS)	88.0	80.5	74.8	82.0	44.7	91.0
Ours (PA)	94.9	92.4	84.0	91.3	57.5	65.4

Table 1: 3D PCK (higher is better) on the MPI-INF-3DHP dataset after training with MPI-INF-3DHP and H3.6M 3D training sets, and MPII and LSP 2D training sets. We clearly outperform all other methods that use a similar combined 2D and 3D training on this benchmark with both indoor and in-the-wild scenes. This holds true for all evaluation protocols (*unscaled* (US), *glob. scaled* (GS), *Procrustes* (PA)).

4.1. Datasets and evaluation metrics

As training data with ground truth 3D pose, we use a combination of the H3.6M training set, as well as both background augmented and unaugmented MPI-INF-3DHP training sets which consist of 350k training images in total. As in-the-wild training images with only 2D pose annotation we use the MPII [1] and LSP [13] [14] datasets which are augmented by randomly cropping, translating and rotating the images.

At test time, we compare against other previously proposed methods on both standard H3.6M and MPI-INF-3DHP test data to show both general 3D pose prediction accuracy, as well as state-of-the-art generalization on outdoor scenes. We also qualitatively visualize the state-of-the-art accuracy of our algorithm on in-the-wild images (see Figure 3).

The quantitative performance is evaluated by comparing the Mean Per Joint Position Error (MPJPE), the Percentage of Correct 3D Keypoints (3D PCK) under 150 mm radius [21], as well as the Area Under Curve (AUC) metric which corresponds to the thresholds of the 3D PCK. Since evaluation protocols in previous work are not uniform, we quantitatively evaluate under the three most commonly used protocols: (i) 3D joint predictions are neither scaled nor aligned to ground truth (*unscaled*), (ii) 3D joint predictions are globally scaled with ground truth scale before evaluation (*glob. scaled*), and (iii) 3D joint predictions are aligned to ground truth with full Procrustes alignment (*Procrustes*). We further follow standard practice by crop-

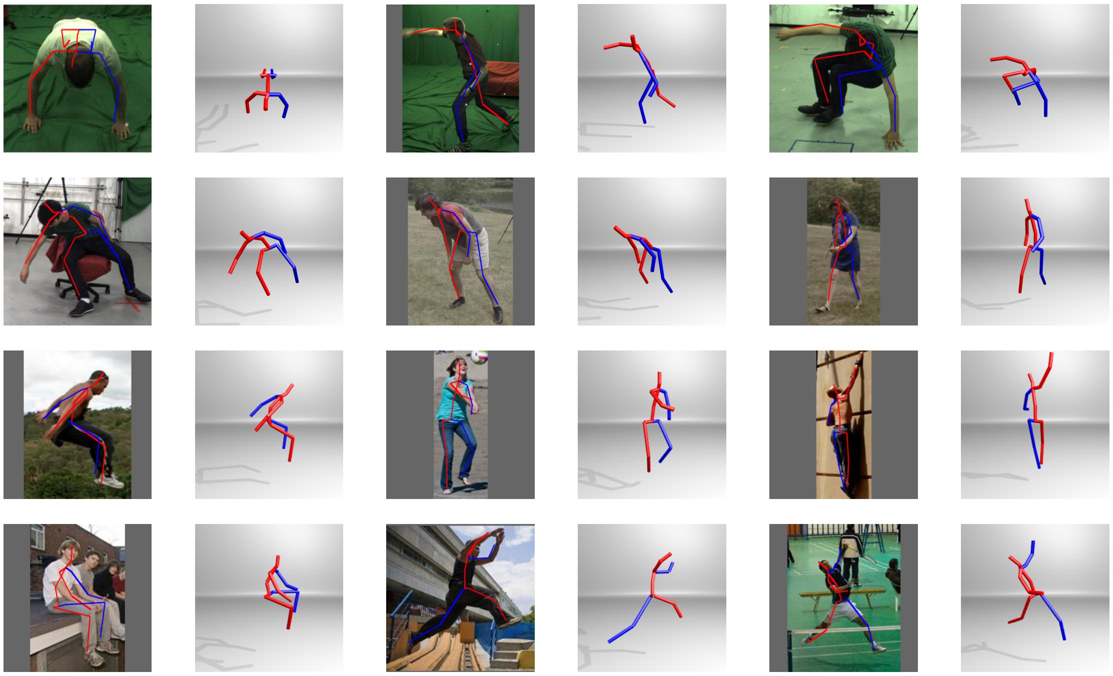

Figure 3: Qualitative examples from the MPI-INF-3DHP test set (first to third rows) and LSP (fourth and fifth rows). Refer to the supplementary document for more examples.

ping a tight bounding box in test images using 2D ground truth information. Since cropping essentially performs a virtual rotation from the original camera, we use perspective correction [21] to re-align the pose to the correct view.

4.2. Training procedure

As outlined earlier, we train our network in two stages. We first pre-train the feature extractor network on the 2D heatmap regression task on both MPII [1] and LSP [14, 13] datasets. At this stage, the network is trained for 186k iterations with a minibatch size of 21. The initial learning rate is 0.05 which we decay exponentially.

After pre-training, we use the learned weights to initialize the weights of the full 3D pose prediction network. The full network is then trained on both the 3D labeled studio data as well as the in-the-wild data with only 2D annotations. Image data with 3D and 2D annotations are both fed into the network with a minibatch size of 10 to train for 240k iterations. For this second stage, we again start the training using a learning rate of 0.05 with a decay over 60k iterations. We use Adadelta with a momentum of 0.9 in both training stages.

We empirically found that using learning rate discrepancy on the pre-trained layers to preserve in-the-wild features, as suggested by [21], is necessary to achieve good generalization if 3D training data is very limited or more biased. We found that a learning rate discrepancy with a factor of 100 when training using H3.6M data as the only source of 3D pose labels yields the best result when tested on the MPI-INF-3DHP dataset. On the other hand, the best performance is achieved without using any such discrepancies when training on both H3.6M and the augmented data of MPI-INF-3DHP as source of 3D labels. This suggest that foreground and background augmentation of the 3D data can further close the domain gap between the indoor and outdoor scenes.

4.3. Quantitative comparison

Table 1 compares our method on the MPI-INF-3DHP benchmark against the closest competing approaches that can be trained on both, images with 2D and 3D annotations. All methods were trained using both H3.6M and augmented and unaugmented MPI-INF-3DHP 3D datasets, and the LSP and MPII 2D datasets. Unless stated otherwise, we used the H80K samples of the H3.6M dataset which consists of around 41K training samples before augmentation. Our algorithm achieves by far the highest accuracy (across all evaluation protocols), yielding 82.0% 3D PCK, 44.7% AUC and 91.0 mm MPJPE overall (using *glob. scaled* for evaluation). We also achieve the state-of-the-art result specifically on the outdoor scenes with 74.8% 3D PCK. Further, the average 3D PCK of 91.3% is the highest ever reported by all algorithms that evaluated on the MPI-INF-3DHP, irrespective of what training data they used. Table 4 further shows the comparison of our approach to other methods on MPI-INF-3DHP, when all methods are trained

	Direction	Discussion	Eating	Greeting	Phoning	Posing	Purchases	Sitting	Sit down	Smoke	Take photo	Waiting	Walk	Walk dog	Walk pair	Average
Mehta* [21]	59.7	69.7	60.6	68.8	76.4	59.1	75.0	96.2	122.9	70.8	85.4	68.5	54.4	82.0	59.8	74.1
Mehta* [22]	62.6	78.1	63.4	72.5	88.3	63.1	74.8	106.6	138.7	78.8	93.8	73.9	55.8	82.0	59.6	80.5
Pavlakos [26]	67.4	72.0	66.7	69.1	72.0	65.0	68.3	83.7	96.5	71.7	77.0	65.8	59.1	74.9	63.2	71.9
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
Zhou* [43]	54.8	60.7	58.2	71.4	62.0	53.8	55.9	75.2	111.6	64.1	65.5	66.1	63.2	51.4	55.3	64.9
Yang* [41]	51.5	58.9	50.4	57.0	62.1	49.8	52.7	69.2	85.2	57.4	65.4	58.4	60.1	43.6	47.7	58.6
Sun* [32]	52.8	54.8	54.2	54.3	61.8	53.1	53.6	71.7	86.7	61.5	67.2	53.4	47.1	61.6	53.4	59.1
Kanazawa* [16]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	88.0
Luvizon* [19]	49.2	51.6	47.6	50.5	51.8	48.5	51.7	61.5	70.9	53.7	60.3	48.9	44.4	57.9	48.9	53.2
Dabral* [7]	46.9	53.8	47.0	52.8	56.9	45.2	48.2	68.0	94.0	55.7	63.6	51.6	40.3	55.4	44.3	55.5
Ours* (H80K)	57.1	69.6	61.6	66.0	73.4	57.1	70.9	89.8	109.2	68.6	81.3	65.8	54.3	78.4	58.2	71.1
Ours* (5 fps)	54.0	65.1	58.5	62.9	67.9	54.0	60.6	82.7	98.2	63.3	75.0	61.2	50.0	66.9	56.5	65.7

Table 2: Mean Per Joint Position Error (MPJPE) on H3.6M when trained on H3.6M (ours are *glob. scaled* for evaluation). (*) indicates methods that also use 2D labeled datasets during training or pre-training.

	Direct.	Discuss	Eat	Greet	Phone	Pose	Purch.	Sit	SitD	Smoke	Photo	Wait	Walk	WalkD	WalkP	Avg.
Sun* [32]	42.1	44.3	45.0	45.4	51.5	43.2	41.3	59.3	73.3	51.0	53.0	44.0	38.3	48.0	44.8	48.3
Kanazawa* [16]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	56.8
Dabral* [7]	32.8	36.8	42.5	38.5	42.4	35.4	34.3	53.6	66.2	46.5	49.0	34.1	30.0	42.3	39.7	42.2
Omran [24]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	59.9
Ours* (H80K)	46.1	51.3	46.8	51.0	55.9	43.9	48.8	65.8	81.6	52.2	59.7	51.1	40.8	54.8	45.2	53.4
Ours* (5 fps)	43.7	46.9	45.4	48.0	50.2	40.6	41.6	60.7	75.6	48.8	54.9	46.8	36.9	47.5	43.9	49.2

Table 3: Mean Per Joint Position Error (MPJPE) on H3.6M when trained on H3.6M. (*) indicates methods that also use 2D labeled datasets during training or pre-training. (*Procrustes* for evaluation).

Method	**PCK**	**AUC**	**MPJPE**
Mehta *et al.* [21]	64.7	31.7	-
Yang *et al.* [41]	69.0	32.0	-
Zhou *et al.* [43]	69.2	32.5	-
Ours (*unscaled*)	69.6	35.5	127.0
Ours (*glob. scaled*)	70.4	36.0	129.1
Ours (*Procrustes*)	82.9	45.4	92.0

Table 4: Comparison on MPI-INF-3DHP after training on H3.6M only. We outperform all other approaches in all metrics and testing protocols.

using only H3.6M as the source of 3D pose labels. Also here, our method achieves the highest accuracy in terms of 3D PCK and AUC on the basis of all three evaluation protocols.

Finally, we also compare our method by only using the H80K samples of H3.6M as the 3D pose dataset and testing on every 64*th* frame of the S9 and S11 subjects in H3.6M, see Table 2 (we use *glob. scaled* following [43][41][7]) and Table 3 (*Procrustes*). On this test set which is heavily biased to in-studio data of a single background our method geared for in-the wild generalization cannot beat the best performing methods. However, it still achieves competitive accuracy. When we increased the number of training data by sampling from H3.6M at 5 frames per second, our

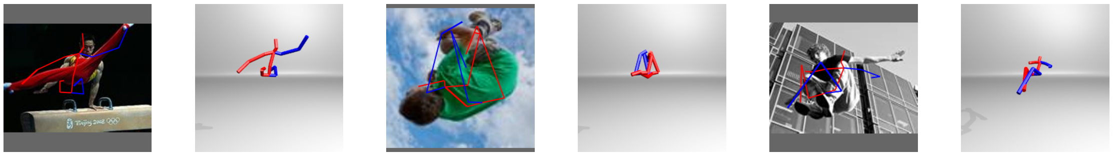

Figure 4: Examples of prediction failures by our proposed method.

method achieved a better MPJPE of 65.7 mm while maintaining competitive result when tested on MPI-INF-3DHP with 71.2% 3D PCK and 36.3% AUC. When using *Procrustes* during comparison, we achieve a state-of-the-art accuracy of 53.4 mm average MPJPE when trained using H80K samples and 49.2 mm average MPJPE when trained using H3.6M data sampled at 5 fps. Notably, here we also outperform other methods that use some form of pose projection operation related to our architecture and regularization with a statistical body model, namely [16] and [24].

4.4. Ablation study

We run an ablation study to measure the effectiveness of our proposed contributions (Table 5). We use a direct 3D pose regression method with 2D pose pre-training without the explicit 2D pose loss in the feature space and without the 2D-from-3D projection loss as baseline. The baseline is trained on 3D data only and uses both joint position and bone losses as training objective. We train all of the comparison results on the H80K samples of H3.6M and then performed the evaluation tests on the MPI-INF-3DHP dataset.

The baseline reaches 62.3% 3D PCK. Using the explicit 2D pose in the latent feature space allows us to use the outdoor data during the training. This addition improves the performance by 3.1% against the baseline. Similarly, adding the 3D-to-2D projection loss improves the performance of the method even without the explicit 2D pose in latent feature space. Using both the proposed components advances the result to the state-of-the-art result with 70.4% 3D PCK.

4.5. Qualitative results and further discussion

We visualize example prediction results on MPI-INF-3DHP and LSP test images in Figure 3. Our method performs consistently well on studio, general indoor and in-the-wild images.

We show several failure cases in Figure 4. Our method can fail on challenging poses which are heavily (self-) occluded, on poses seen from unusual camera angles, or poses which are from what was seen in the training set. Such failure cases are common to many monocular 3D pose estimation approaches. The supplementary document shows additional failure examples of our method.

Method	**PCK**	**AUC**
Baseline (direct 3D prediction + bone loss)	62.3	30.3
+ 2D latent loss + outdoor data	66.4	33.0
+ 3D-to-2D projection + outdoor data	69.5	35.3
+ 2D latent loss + outdoor data + 3D-to-2D projection	70.4	36.0

Table 5: Ablation study on MPI-INF-3DHP test data (split into scene sub-categories: in-studio with green screen (GS), and more in-the-wild scenes indoors (No GS) and outdoors (Outdoor)). Only H3.6M data with ground truth 3D labels were used for training. 3D predictions are globally scaled.

5. Conclusion

We proposed a new deep learning architecture for 3D human pose estimation from monocular color images. It is designed for training on both, more scarcely available real images with ground truth 3D pose labels, and more widely available in-the-wild images with only 2D pose labels. Our architecture augments a backbone 3D pose inference network with an explicit disentangled 2D pose representation in latent feature space and a learned 3D-to-2D projection model. Our algorithm achieves state-of-the-art performance on the in-studio H3.6M dataset and clearly outperforms related work on the more challenging MPI-INF-3DHP benchmark with in-the-wild images.

6. Acknowledgments

This work was supported by the ERC Consolidator Grant 4DRepLy (770784). Gerard Pons-Moll is funded by the Deutsche Forschungsgemeinschaft (DFG. German Research Foundation) - 409792180.

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 5, 6

[2] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL: Automatic estimation of 3D

human pose and shape from a single image. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pages 561–578. Springer International Publishing, 2016. 2, 3

[3] E. Brau and H. Jiang. 3d human pose estimation via deep learning from 2d annotations. In *Fourth International Conference on 3D Vision, 3DV 2016, Stanford, CA, USA, October 25-28, 2016*, pages 582–591, 2016. 3

[4] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 2

[5] C. Chen and D. Ramanan. 3d human pose estimation = 2d pose estimation + matching. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5759–5767, 2017. 2, 3

[6] W. Chen, H. Wang, Y. Li, H. Su, Z. Wang, C. Tu, D. Lischinski, D. Cohen-Or, and B. Chen. Synthesizing training images for boosting human 3d pose estimation. In *3D Vision (3DV)*, 2016. 1

[7] R. Dabral, A. Mundhada, U. Kusupati, S. Afaque, A. Sharma, and A. Jain. Learning 3d human pose from structure and motion. 2018. 2, 3, 5, 7

[8] A. Elhayek, E. Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt. Efficient convnet-based marker-less motion capture in general scenes with a low number of cameras. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[10] Y. Huang, M. Kaufmann, E. Aksan, M. J. Black, O. Hilliges, and G. Pons-Moll. Deep inertial poser learning to reconstruct human pose from sparseinertial measurements in real time. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 37(6):185:1–185:15, nov 2018. 1

[11] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. October 2016. 2

[12] C. Ionescu, D. Papava, V. Olaru, and C. 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. 2, 5

[13] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proc. BMVC*, pages 12.1–11, 2010. doi:10.5244/C.24.12. 5, 6

[14] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR 2011*, 2011. 5, 6

[15] H. Joo, T. Simon, and Y. Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8320–8329, 2018. 1

[16] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 3, 7, 8

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. 2012. 1

[18] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 3

[19] D. Luvizon, D. Picard, and H. Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *CVPR*, 2018. 7

[20] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings IEEE International Conference on Computer Vision (ICCV)*, Piscataway, NJ, USA, Oct. 2017. IEEE. 2, 3, 4, 7

[21] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017. 1, 2, 5, 6, 7

[22] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics*, 36(4), 2017. 1, 2, 5, 7

[23] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 2

[24] M. Omran, C. Lassner, G. Pons-Moll, P. V. Gehler, and B. Schiele. Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. In *3DV*, Sept. 2018. 3, 7, 8

[25] G. Pavlakos, X. Zhou, and K. Daniilidis. Ordinal depth supervision for 3D human pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 3, 4

[26] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 7

[27] G. Pons-Moll, D. J. Fleet, and B. Rosenhahn. Posebits for monocular human pose estimation. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2345–2352, Columbus, Ohio, USA, jun 2014. 3

[28] H. Rhodin, N. Robertini, D. Casas, C. Richardt, H.-P. Seidel, and C. Theobalt. General automatic human shape and motion capture using volumetric contour cues. In *European conference on computer vision*, pages 509–526. Springer, 2016. 1

[29] H. Rhodin, M. Salzmann, and P. Fua. Unsupervised geometry-aware representation learning for 3d human pose estimation. 2018. 3

[30] H. Rhodin, J. Sprri, I. Katircioglu, V. Constantin, F. Meyer, E. Mller, M. Salzmann, and P. Fua. Learning monocular 3d human pose estimation from multi-view images. page 16, 2018. 2

[31] C. Stoll, N. Hasler, J. Gall, H.-P. Seidel, and C. Theobalt. Fast articulated motion tracking using a sums of gaussians body model. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 951–958, Washington, DC, USA, 2011. IEEE Computer Society. 1

[32] X. Sun, J. Shang, S. Liang, and Y. Wei. Compositional human pose regression. 2017. 2, 4, 7

[33] B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, and P. Fua. Structured prediction of 3d human pose with deep neural networks. In *British Machine Vision Conference*, 2016. 2

[34] D. Tomè, C. Russell, and L. Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5689–5698, 2017. 2, 3

[35] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. Efficient object localization using convolutional networks. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 648–656, 2015. 2

[36] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, 2014. 2

[37] T. von Marcard, R. Henschel, M. Black, B. Rosenhahn, and G. Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, sep 2018. 1

[38] T. von Marcard, B. Rosenhahn, M. Black, and G. Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. *Computer Graphics Forum 36(2), Proceedings of the 38th Annual Conference of the European Association for Computer Graphics (Eurographics)*, pages 349–360, 2017. 1

[39] C. Wang, Y. Wang, Z. Lin, and A. Yuille. Robust 3d human pose estimation from single images or video sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2018. 2

[40] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4732, 2016. 2

[41] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, and X. Wang. 3d human pose estimation in the wild by adversarial learning. In *CVPR*, 2018. 1, 2, 3, 7

[42] H. Yasin, U. Iqbal, B. Krüger, A. Weber, and J. Gall. A dual-source approach for 3d pose estimation from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition 2016 (CVPR)*, June 2016. 2

[43] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei. Towards 3d human pose estimation in the wild: A weakly-supervised approach. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1, 2, 3, 7

[44] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei. Deep kinematic pose regression. In *ECCV Workshop on Geometry Meets Deep Learning*, 2016. 2