Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose

Georgios Pavlakos[1], Xiaowei Zhou[1], Konstantinos G. Derpanis[2], Kostas Daniilidis[1]
[1] University of Pennsylvania   [2] Ryerson University

Abstract

This paper addresses the challenge of 3D human pose estimation from a single color image. Despite the general success of the end-to-end learning paradigm, top performing approaches employ a two-step solution consisting of a Convolutional Network (ConvNet) for 2D joint localization and a subsequent optimization step to recover 3D pose. In this paper, we identify the representation of 3D pose as a critical issue with current ConvNet approaches and make two important contributions towards validating the value of end-to-end learning for this task. First, we propose a fine discretization of the 3D space around the subject and train a ConvNet to predict per voxel likelihoods for each joint. This creates a natural representation for 3D pose and greatly improves performance over the direct regression of joint coordinates. Second, to further improve upon initial estimates, we employ a coarse-to-fine prediction scheme. This step addresses the large dimensionality increase and enables iterative refinement and repeated processing of the image features. The proposed approach outperforms all state-of-the-art methods on standard benchmarks achieving a relative error reduction greater than 30% on average. Additionally, we investigate using our volumetric representation in a related architecture which is suboptimal compared to our end-to-end approach, but is of practical interest, since it enables training when no image with corresponding 3D groundtruth is available, and allows us to present compelling results for in-the-wild images.

1. Introduction

Estimating the full-body 3D pose of a human from a single monocular image is an open challenge, which has garnered significant attention since the early days of computer vision [18]. Given its ill-posed nature, researchers have generally approached 3D human pose estimation in simplified settings, such as assuming background subtraction is feasible [1], relying on groundtruth 2D joint locations to estimate 3D pose [26, 43], employing additional camera views [7, 15], and capitalizing on temporal consistency to improve upon single frame predictions [38, 3]. This di-
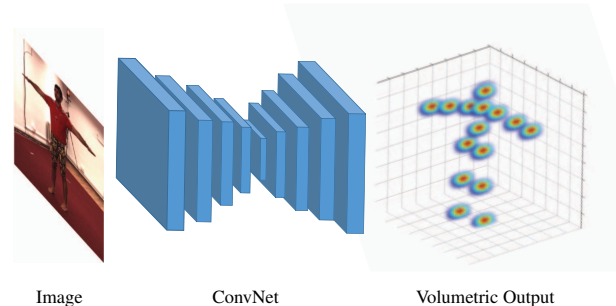

Figure 1: Illustration of our volumetric representation for 3D human pose. We discretize the space around the subject and use a ConvNet to predict per voxel likelihoods for each joint from a single color image.

versity of assumptions and additional information sources exemplifies the challenge presented by the task.

With the introduction of more powerful discriminative approaches, such as Convolutional Networks (ConvNets), many of these restrictive assumptions have been relaxed. End-to-end learning approaches attempt to estimate 3D pose directly from a single image by addressing it as coordinate regression [19, 35], nearest neighbor between images and poses [20], or classification over a set of pose classes [27]. Yet to date, these approaches have been outperformed by more traditional two-step pipelines, e.g., [45, 6]. In these cases, ConvNets are used only for 2D joint localization and 3D poses are generated during a post-processing optimization step. Combining accurate 2D joint localization with strong and expressive 3D priors has been proven to be very effective. In this work, we show that ConvNets are able to provide much richer information than simply 2D joint locations.

To fully exploit the potential of ConvNets in the context of 3D human pose, we propose the following items, and justify them empirically. First, we cast 3D pose estimation as a keypoint localization problem in a discretized 3D space. Instead of directly regressing the coordinates of the joints (e.g., [19, 35]), we train a ConvNet to predict per voxel likelihoods for each joint in this volume. This volumetric representation, illustrated in Figure 1, is much more sensible for the 3D nature of our problem and improves learning.

Effectively, for every joint, the volumetric supervision provides the network with groundtruth for each voxel in the 3D space. This provides much richer information than a set of world coordinates. The empirical results also validate the superiority of our proposed form of supervision.

Second, to deal with the increased dimensionality of the volumetric representation, we propose a coarse-to-fine prediction scheme. As demonstrated in the 2D pose case, intermediate supervision and iterative estimation are particularly effective strategies [40, 8, 21]. For our volumetric representation though, naively stacking an increasing number of components and refining the estimates is not an effective solution, as shown empirically. Instead, we gradually increase the resolution of the supervision volume for the most challenging z-dimension (depth), during the processing. This coarse-to-fine supervision, illustrated schematically in Figure 2, allows for more accurate estimates after each step. We empirically demonstrate the advantage of this practice over naively stacking more components together.

Our proposed approach achieves state-of-the-art results on standard benchmarks, outperforming both ConvNet-only and hybrid approaches that post-process the 2D output of a ConvNet. Additionally, we investigate using our volumetric representation within a related architecture that decouples 2D joint localization and 3D joint reconstruction. In particular, we use two separate networks (the output of one serves as the input to the other) and two non-corresponding data sources, i.e., 2D labeled imagery to train the first component and an independent 3D data source (e.g., MoCap) to train the second one separately. While this architecture has practical benefits (e.g., predicting 3D pose for in-the-wild images), we show empirically that it underperforms compared to our end-to-end approach when images with corresponding 3D groundtruth are available for training. This finding further underlines the benefit of predicting 3D pose directly from an image, whenever this is possible, instead of using 2D joint localization as an intermediate step.

In summary, we make the four following contributions:

- we are the first to cast 3D human pose estimation as a 3D keypoint localization problem in a voxel space using the end-to-end learning paradigm;

- we propose a coarse-to-fine prediction scheme to deal with the large dimensionality of our representation and enable iterative processing to realize further benefits;

- our proposed approach achieves state-of-the-art results on standard benchmarks, surpassing both ConvNet-only and hybrid approaches that employ ConvNets for 2D pose estimation, with a relative error reduction that exceeds 30% on average;

- we show the practical use of our volumetric representation in cases when end-to-end training is not an option and present compelling results on in-the-wild images.

2. Related work

The literature on 3D human pose estimation is vast with approaches addressing the problem in a variety of settings. Here, we survey works that are most relevant to ours with a focus on ConvNet-based approaches; we refer the reader to a recent survey [29] for a more complete literature review.

The majority of recent ConvNet-only approaches cast 3D pose estimation as a coordinate regression task, with the target output being the spatial x, y, z coordinates of the human joints with respect to a known root joint, such as the pelvis. Li and Chan [19] pretrain their network with maps for 2D joint classification. Tekin *et al.* [35] include a pretrained autoencoder within the network to enforce structural constraints on the output. Ghezelghieh *et al.* [13] employ viewpoint prediction as a side task to provide the network with global joint configuration information. Zhou *et al.* [44] embed a kinematic model to guarantee the validity of the regressed pose. Park *et al.* [22] concatenate the 2D joint predictions with image features to improve 3D joint localization. Tekin *et al.* [36] include temporal information in the joint predictions by extracting spatiotemporal features from a sequence of frames. In contrast to all these approaches, we adopt a volumetric representation of the human pose, and regress the per voxel likelihood for each joint separately. This proves to have significant advantages for the network performance and provides a richer output compared to the low-dimensional vector of joint coordinates.

An alternative approach to the classical regression paradigm is proposed by Li *et al.* [20]. During training, they learn a common embedding between color images and 3D poses. At test time, the test image is coupled with each candidate pose and forwarded through the network; the input image is assigned the candidate pose with the maximum network score. This is a form of nearest neighbor classification which is highly inefficient due to the requirement of multiple forward network passes. On the other hand, Rogez and Schmid [27] cast pose estimation as a classification problem. Given a predefined set of pose classes, each image is assigned to the class with the highest score. This guarantees a valid global pose prediction, but the approach is constrained by the poses in the original classes and thus returns only a rough pose estimate. In contrast to the inefficient nearest neighbor approach and the coarse classification approach, our volume regression allows for much more accurate 3D joint localization, while also being efficient.

Despite the interest in end-to-end learning, ConvNet-only approaches underperform those that employ a ConvNet for the 2D localization of joints, and produce 3D pose with a subsequent optimization step. Zhou *et al.* [45] utilize a standard 2D pose ConvNet to localize the joints and retrieve the 3D pose using an optimization scheme over a sequence of monocular images. Similarly, Du *et al.* [10] include height-maps of the human body to improve 2D joint
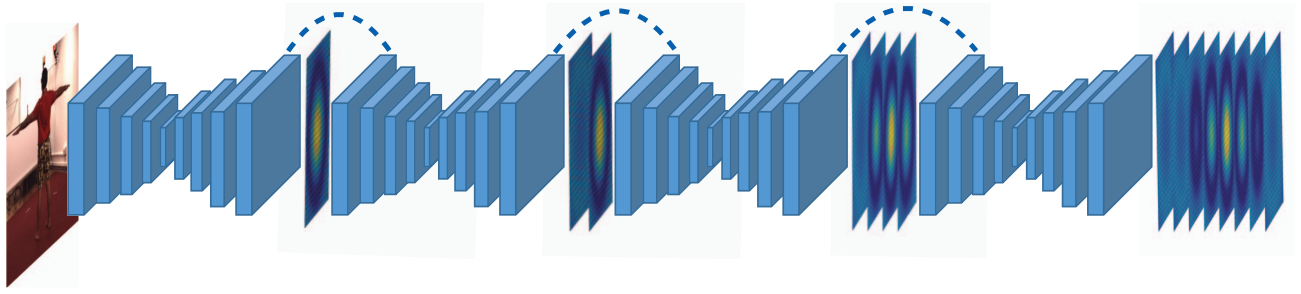
Figure 2: Illustration of our coarse-to-fine volumetric approach for 3D human pose estimation from a single image. The input is a single color image and the output is a dense 3D volume with separate per voxel likelihoods for each joint. The network consists of multiple fully convolutional components [21], which are supervised in a coarse-to-fine fashion, to deal with the large dimensionality of our representation. 3D heatmaps are synthesized for supervision by increasing the resolution for the most challenging z-dimension (depth) after each component. The dashed lines indicate that the intermediate heatmaps are fused with image features to produce the input for the next fully convolutional component. For presentation simplicity, the illustrated heatmaps correspond to the location of only one joint.

localization. Bogo *et al.* [6] use the joints predicted by a 2D ConvNet and fit a statistical body shape model to recover the full shape of the human body. In contrast, our approach achieves state-of-the-art results with a single network. Furthermore, it provides a rich 3D output, amenable to post-processing, such as pictorial structures optimization to constrain limb lengths, or temporal filtering.

Another issue that has been addressed in the context of using ConvNets for 3D human pose is the scarcity of training data. Chen *et al.* [9] use a graphics renderer to create images with known groundtruth. Similarly, Ghezelghieh *et al.* [13] augment the training set with synthesized examples. A collage approach is proposed by Rogez and Schmid [27], where parts from in-the-wild images are combined to create additional images with known 3D poses. However, there is no guarantee that the statistics of the synthetic examples match those of real images. To investigate the data scarcity issue, we take inspiration from the 3D Interpreter Network [41], which decouples the 3D pose estimation task into 2D localization and 3D reconstruction within a single ConvNet. In contrast, rather than using a predefined linear basis for 3D reconstruction, we predict 3D joint locations directly with our volumetric representation. This demonstrates the practical use of our volumetric representation even when end-to-end training is not an option.

Finally, while we do not compare explicitly with multi-view pose estimation work (e.g., [12, 31, 4, 11]), it is interesting to note that the representation of 3D human pose in a discretized 3D space has also been previously adopted in multi-view settings [7, 15, 23], where it was used to accommodate predictions from different viewpoints. For single view pose estimation, it has been considered in the context of random forests [16]. This approach suffered from large execution time (around three minutes), and required an additional refinement step using a pictorial structures model.

In stark contrast, our network can provide complete volume predictions with a single forward pass in a few milliseconds, needs no additional refinement (although it is still a possibility) to provide state-of-the-art results, and is integrated within a coarse-to-fine prediction scheme to deal with excessive dimensionality.

3. Technical approach

The following subsections summarize our technical approach. Section 3.1 describes the proposed volumetric representation for 3D human pose and discusses its merits. Next, Section 3.2 describes our coarse-to-fine prediction approach that addresses the high dimensional nature of our output representation. Finally, Section 3.3 describes the use of our volumetric representation within a related decoupled architecture and discusses its relative merits compared to our coarse-to-fine volumetric prediction approach.

3.1. Volumetric representation for 3D human pose

The problem of 3D human pose estimation using ConvNets has been primarily approached as a coordinate regression problem. In this case, the target of the network is a $3N$-dimensional vector comprised of the concatenation of the x, y, z coordinates of the N joints of the human body. For training, an $\mathcal{L}_2$ regression loss is employed:

$$\mathcal{L} = \sum_{n=1}^{N} \|\boldsymbol{x}_{gt}^n - \boldsymbol{x}_{pr}^n\|_2^2, \tag{1}$$

where $\boldsymbol{x}_{gt}^n$ is the groundtruth and $\boldsymbol{x}_{pr}^n$ is the predicted location for joint n. The location of each joint is expressed globally, with respect to a root joint, or locally, with respect to its parent joint in the kinematic tree. The second formulation has some benefits, as discussed also by Li *et*

al. [19] (e.g., easier to learn to predict small, local deviations), but still suffers from the fact that small errors can easily propagate hierarchically to children joints of the kinematic tree. In general, despite its simplicity, the coordinate regression approach makes the problem highly non-linear and presents problems for the learning procedure. These issues have previously been identified in the context of 2D human pose [37, 24].

To improve learning, we propose a volumetric representation for 3D human pose. The volume around the subject is discretized uniformly in each dimension. For each joint we create a volume of size $w \times h \times d$. Let $p^n_{(i,j,k)}$ denote the predicted likelihood of joint n being in voxel (i, j, k). To train this network, the supervision is also provided in volumetric form. The target for each joint is a volume with a 3D Gaussian centered around the groundtruth position $\boldsymbol{x}^n_{gt} = (x, y, z)$ of the joint in the 3D grid:

$$G_{i,j,k}(\boldsymbol{x}^n_{gt}) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x-i)^2+(y-j)^2+(z-k)^2}{2\sigma^2}}, \quad (2)$$

where the value $\sigma = 2$ is used for our experiments. For training, we use the mean squared error loss:

$$\mathcal{L} = \sum_{n=1}^{N} \sum_{i,j,k} \|G_{(i,j,k)}(\boldsymbol{x}^n_{gt}) - p^n_{(i,j,k)}\|^2. \quad (3)$$

In theory, the output of the network is four dimensional, i.e., $(w \times h \times d \times N)$, but in practice we organize it in channels, thus our output is three dimensional, i.e., $w \times h \times dN$. The voxel with the maximum response in each 3D grid is selected as the joint's 3D location.

A major advantage of the volumetric representation is that it casts the highly non-linear problem of direct 3D coordinate regression to a more manageable form of prediction in a discretized space. In this case, the predictions do not necessarily commit to a unique location for each joint, but instead an estimate of the confidence is provided for each voxel. This makes it easier for the network to learn the target mapping. A similar argument has been previously put forth in the 2D pose case, validating the benefit of predicting per pixel likelihoods instead of pixel coordinates [37, 24]. In terms of the network architecture, an important benefit of the volumetric representation is that it enables the use of a fully convolutional network for prediction. Here, we adopt the hourglass design [21]. This leads to less network parameters than using fully connected layers for coordinate regression or pose classification. Finally, in terms of the predicted output, besides being more accurate, our network predictions in the form of dense 3D heatmaps are useful for subsequent post-processing applications. For example, structural constraints can be enforced with the use of a 3D Pictorial Structures model, e.g., [7, 23]. Another option is to use the dense predictions in a filtering framework in cases where multiple input frames are available.

3.2. Coarse-to-fine prediction

A design choice that has been particularly effective in the case of 2D human pose is the iterative processing of the network output [8, 40, 21]. Instead of using a single component with a single output, the network is forced to produce predictions in multiple processing stages. These intermediate predictions are gradually refined to produce more accurate estimates. Additionally, the use of intermediate supervision on the "earlier" outputs allows for a richer gradient signal, which has been demonstrated empirically as an effective learning practice [17, 34].

Inspired by the success of iterative refinement in the context of 2D pose, we also consider a gradual refinement scheme. Empirically, we found that naively stacking multiple components yielded diminishing returns because of the large dimensionality of our representation. In fact, for the highest 3D resolution of $64 \times 64 \times 64$ with 16 joints, we would need to estimate the likelihood for more that 4 million voxels. To deal with this curse of dimensionality, we propose to use a coarse-to-fine prediction scheme. In particular, the first steps are supervised with lower resolution targets for the (most challenging and technically unobserved) z-dimension. Precisely, we use targets of size $64 \times 64 \times d$ per joint, where d typically takes values from the set $\{1, 2, 4, 8, 16, 32, 64\}$. An illustration of this supervision approach is given in Figure 2.

This strategy makes training more effective, and allows us to benefit from stacking multiple components together without suffering from overfitting or dimensionality issues. Intuitively, easier versions of the task are presented to the network during the early stages of processing, and the complexity increases gradually. This postpones the harder decisions until the very end of the processing, when all the available information has been processed and consolidated.

3.3. Decoupled architecture with volumetric target

To further show the versatility of the proposed volumetric representation, we also employ it in a scenario where end-to-end training is not an option. This is usually the case for in-the-wild images, where accurate, large-scale acquisition of 3D groundtruth is not feasible. Inspired by the 3D Interpreter Network [41], we decouple 3D pose estimation in two sequential steps consisting of predicting 2D keypoint heatmaps, followed by an inference step of the 3D joint positions with our volumetric representation. The first step can be trained with 2D labeled in-the-wild imagery, while the second step requires only 3D data (e.g., MoCap). Independently, each of these sources are abundantly available.

This training strategy is useful for practical scenarios, and we present compelling results for in-the-wild images (Sec. 4.6). However, it remains suboptimal compared to our end-to-end approach when images with corresponding 3D groundtruth are available for training. Figure 3 provides an

(a) Decoupled architecture
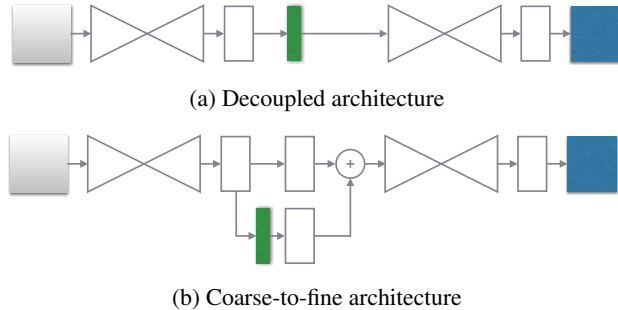

(b) Coarse-to-fine architecture

Figure 3: Schematic comparison of a decoupled architecture versus our coarse-to-fine architecture with intermediate supervision at the coarsest level (2D heatmaps). Blue blocks indicate 3D heatmaps, while green blocks indicate 2D heatmaps. **Decoupled architecture**: The 2D heatmaps are provided directly as input to the second part of the network, which effectively operates as a 2D-to-3D reconstruction component. Note, no image features are processed in the second component, only information about 2D joint locations. **Coarse-to-fine architecture**: We use 2D heatmaps as intermediate supervision, which are then combined with image features, effectively carrying information both from the image and the 2D locations of the joints.

illustration of each architecture in a simplified setting with two hourglasses. It can be seen that the decoupled case is related to our course-to-fine architecture when the resolution of the intermediate supervision is set to $d = 1$ resulting in 2D heatmaps. A crucial difference between the two architectures is that our coarse-to-fine approach combines the produced 2D heatmaps with intermediate image features. This way, the rest of the network can process information both about the image and the 2D joint locations. On the other hand, a decoupled network processes the 2D heatmaps directly and attempts to reconstruct 3D locations without further aid by image-based evidence. In cases where the heatmaps are grossly erroneous, the 3D predictions can be lead astray. In Sec. 4.4, we show empirically that when images with corresponding 3D groundtruth are available, our coarse-to-fine architecture outperforms the decoupled one.

4. Empirical evaluation

4.1. Datasets

We present extensive quantitative evaluation of our coarse-to-fine volumetric approach on three standard benchmarks for 3D human pose: Human3.6M [14], HumanEva-I [30] and KTH Football II [15]. Additionally, qualitative results are presented on the MPII human pose dataset [2], since no 3D groundtruth is available.

Human3.6M: It contains video of 11 subjects performing a variety of actions, such as "Walking", "Sitting" and "Phoning". We follow the same evaluation protocol as prior work [20, 45]. In particular, Subjects S1, S5, S6, S7 and S8 were used for training, while subjects S9 and S11 were used for testing. The original videos were downsampled from 50fps to 10fps We employed all camera views and trained a single model for *all actions*, instead of training action-specific models [20, 45].

HumanEva-I: It is a smaller dataset compared to Human3.6M, with fewer subjects and actions. Following the standard protocol [16, 42], we evaluated on "Walking" and "Jogging" from subjects S1, S2 and S3. The training sequences of these subjects and actions were used for training, and the corresponding validation sequences for testing. As done with the Human3.6M evaluation, we train a single model using the frames for all users and actions.

KTH Football II: The images are taken from a professional football match, and 3D groundtruth is provided only for a very small number of them. The limited available groundtruth is not very accurate, since it was generated by combining manual 2D annotations from multiple views. In this case, image-to-3D training is not a practical option. Instead, we report results using our volumetric representation within the decoupled architecture described in Sec. 3.3. More specifically, we train the first network component (image to 2D heatmaps) using images from this dataset which provide 2D groundtruth. For the second network component (2D heatmaps to 3D heatmaps), we use all the training MoCap data from Human3.6M dataset. As others [7, 36], we report results using "Sequence 1" from "Player 2" and frames taken from "Camera 1".

MPII: It is a large scale 2D pose dataset containing in-the-wild imagery. It provides 2D annotations but no 3D groundtruth. Like KTH, direct image-to-3D training is not a practical option with this dataset. Instead, we use the decoupled architecture with our volumetric representation. Since we cannot quantify the performance here, we only provide qualitative results.

4.2. Evaluation metrics

For Human3.6M, most approaches report the *per joint 3D error*, which is the average Euclidean distance of the estimated joints to the groundtruth. This is done after aligning the root joints (here the pelvis) of the estimated and groundtruth 3D pose. An alternative metric, which is used by some methods to report results on Human3.6M and HumanEva-I is the *reconstruction error*. It is defined as the per joint 3D error up to a similarity transformation. Effectively, the estimated 3D pose is aligned to the groundtruth by the Procrustes method. Finally, for KTH the percentage of the correctly estimated parts in 3D (3D PCP [7]) is reported. Again, the root joints (here we use the center of the chest) are aligned to resolve the depth ambiguity.

4.3. Implementation details

In our volumetric space, the x-y grid is a uniform discretization within the 2D bounding box in the image and the z grid is a uniform discretization in $[-1, 1]$ (meters) centered at the root joint. This means that we predict the image coordinates for each joint and its metric depth relative to the root. Given the depth of the root joint we can recover the absolute depth for each joint and its metric coordinates in the x-y dimensions. For the component analysis (Section 4.4) we use the groundtruth depth of the root joint, while for the comparison to other methods (Section 4.5) we estimate this depth based on the average skeleton size of each dataset. More details are provided in the supplementary material.

In terms of the architecture, the fully convolutional components of our network, illustrated in Figure 2, are based on the hourglass design [21]. We use the publicly available code to faithfully replicate the architecture. Similarly, we adopt the same training practices, employing rotation augmentation ($\pm 30^o$), scaling augmentation (0.75-1.25), left-right flipping, and using rmsprop for optimization with the batch size equal to 4 and the learning rate set to $2.5\mathrm{e}{-4}$.

Regarding training on individual datasets, for Human3.6M, the network models are trained from scratch, typically for four epochs (approximately 310k iterations). For HumanEva-I, the model is trained from scratch for 120 epochs (approximately 235k iterations), because of the significantly smaller training set size. Finally, for the 2D joint localization network on KTH, we use the publicly available stacked hourglass model trained on MPII [21] and fine-tune it for 20 epochs (approximately 30k iterations).

4.4. Component evaluation

To evaluate the components of our approach, we use Human3.6M to report results, since it is the most complete available benchmark.

Volumetric representation: Our first goal is to demonstrate that regression in a discretized space provides great benefit over coordinate regression. Both versions are implemented with the simplest setting of one hourglass. The only difference between the architectures is that the network for the volumetric predictions is fully convolutional, while for coordinate regression we use a fully connected layer at the end. The results are presented in Table 1. The error of 112.41mm for coordinate regression is comparable to recent reported results with a coordinate regression target output [19, 35, 22, 44]. In comparison, a significant decrease in the error is observed, down to 85.82mm at the highest depth resolution, by using the volumetric output target.

Coarse-to-fine prediction: The next significant improvement to our network is provided by iterative processing of the image features. The naive way to achieve this is by stacking together multiple hourglasses. This is helpful, but

	Average
Coordinate Regression	112.41
Volume Regression ($d = 32$)	92.23
Volume Regression ($d = 64$)	**85.82**

Table 1: Coordinate versus volume regression on Human3.6M. The mean per joint error (mm) across all actions and subjects in the test set are shown.

Naive Stacking					Coarse-to-Fine				
L1	L2	L3	L4	Avg.	L1	L2	L3	L4	Avg.
64	64			80.14	1	64			69.77
					8	64			77.52
64	64	64		78.17	1	2	64		68.49
					1	4	64		72.02
64	64	64	64	75.06	1	2	4	64	64.76

Table 2: Comparison of the Naive Stacking (left) versus Coarse-to-Fine (right) approaches on Human3.6M. The column Li denotes the z-dimension resolution for the supervision provided at the i-th hourglass component (empty if the network has less than i components). We report mean per joint errors (mm) following the standard protocol.

offers only diminishing returns, as is demonstrated in Table 2 (Naive stacking). Instead, our coarse-to-fine supervision approach outperforms the naive one when we use two, three, or four hourglasses (Table 2, Coarse-to-Fine). In fact, our coarse-to-fine version with two hourglasses (69.77mm error) outperforms the naive stacking network with four hourglasses (75.06mm error), despite using less than half of the parameters compared to the deeper network.

Decoupled network with volumetric representation: We investigate the use of a decoupled network combined with our volumetric representation, as described in Sec. 3.3. Our goal is to demonstrate the benefit of predicting the 3D pose directly from image features versus using 2D locations as an intermediate representation. We refer back to Fig. 3 for a schematic representation of the two relevant networks. We train both networks end-to-end to evaluate the architecture performance rather than the benefit of end-to-end training. (In fact we observed that training the two components of the decoupled network independently leads to inferior performance.) Comparative results are provided in Table 3. We present the average across all actions, as well as the six actions with the largest difference between the two networks. Besides being more accurate for every action and on average, our coarse-to-fine approach shows significant improvement for more challenging actions, such as "Sitting" or "Sitting Down". In these cases, 2D joint localization often fails because of severe self-occlusions, providing the second subnetwork with inaccurate 2D heatmaps. Unless we process

Action	Decoupled	Coarse-to-Fine
Phoning	75.00	**66.50**
Sitting	95.25	**76.99**
SittingDown	129.97	**103.67**
Smoking	75.58	**66.99**
Walking	70.17	**59.12**
WalkTogether	76.01	**62.28**
Average	78.10	**69.77**

Table 3: Comparison of our coarse-to-fine network using 2D heatmaps for intermediate supervision (Coarse-to-Fine) versus a decoupled network with a volumetric representation (Decoupled). The reported results are for the six classes of Human3.6M with the largest difference between the two approaches, as well as the average across all actions.

	Sanzari *et al.* [28]	Rogez *et al.* [27]	Bogo *et al.* [6]	Ours
Average	93.2	87.3	82.3	**51.9**

Table 5: Quantitative comparison on Human3.6M among approaches that report reconstruction error (mm). Baseline numbers are taken from the respective papers.

image features as well, 3D localization is bound to fail. This demonstrates the benefit of using information directly from the image for 3D localization versus decoupling the process.

4.5. Comparison with state-of-the-art

Human3.6M: We compare the performance of our approach with previously reported results on Human3.6M. Table 4 presents the mean per joint 3D error results. Note that some previous works [36, 45, 10] leverage a sequence of frames for pose prediction rather than a single frame as considered by our approach. Nonetheless, our network achieves state-of-the-art results across the vast majority of actions and outperforms all other methods on average. Since some works use reconstruction error to report results, we also evaluate using this metric in Table 5. Again, our approach outperforms the other baselines by large margins.

HumanEva-I: Our empirical results for HumanEva-I are presented in Table 6, along with reported results from state-of-the-art approaches. Similar to Human3.6M, our network outperforms all other published approaches.

KTH Football II: We report results for our approach on this dataset in Table 7, comparing with relevant methods. Note, Tekin *et al.* [36] use video for prediction instead of a single frame, while Burenius *et al.* [7] is a multi-view method. Despite this, we outperform the single view approaches and we are competitive with the multi-view results.

4.6. Qualitative results

Figure 4 presents qualitative results for images taken from the aforementioned datasets. Additionally, we demon-

	Walking			Jogging			Avg
	S1	S2	S3	S1	S2	S3	
Radwan *et al.* [25]	75.1	99.8	93.8	79.2	89.8	99.4	89.5
Wang *et al.* [39]	71.9	75.7	85.3	62.6	77.7	54.4	71.3
Simo-Serra *et al.* [32]	65.1	48.6	73.5	74.2	46.6	32.2	56.7
Bo *et al.* [5]	46.4	30.3	64.9	64.5	48.0	38.2	48.7
Kostrikov *et al.* [16]	44.0	30.9	41.7	57.2	35.0	33.3	40.3
Yasin *et al.* [42]	35.8	32.4	41.6	46.6	41.4	35.4	38.9
Ours	**22.3**	**19.5**	**29.7**	**28.9**	**21.9**	**23.8**	**24.3**

Table 6: Quantitative results on HumanEva-I. The numbers are the mean reconstruction errors (mm). Baseline numbers are taken from the respective papers.

	[7]	[7]	[7]	[36]	Ours
Cameras	3	2	1	1 (video)	1
Upper Arms	60	53	14	74	**96**
Lower Arms	35	28	06	49	**83**
Upper Legs	100	88	63	**98**	**98**
Lower Legs	90	82	41	77	**88**

Table 7: Quantitative results on KTH Football II. The numbers are the mean PCP scores (the higher the better). Baseline numbers are taken from the respective papers. We indicate how many cameras each approach uses, and highlight the best performance for single view approaches.

strate 3D reconstructions on MPII which offers greater visual variety due to its in-the-wild nature. Despite the challenging poses present in the MPII examples, our volumetric representation produces compelling 3D predictions.

5. Summary

Our paper addressed the challenging problem of 3D human pose estimation from a single color image. Departing from recent ConvNet approaches, we cast the problem as 3D keypoint localization in a discretized space around the subject. We integrated this volumetric representation with a coarse-to-fine supervision scheme to deal with the high dimensionality and enable iterative processing. We demonstrated that our contributions were crucial to achieve state-of-the-art results on the standard benchmarks with a relative error reduction greater than 30% on average. Furthermore, we used our volumetric representation within a decoupled architecture, making it of practical use for in-the-wild images even when end-to-end training is not feasible.

Project Page: https://www.seas.upenn.edu/~pavlakos/projects/volumetric

	Directions	Discussion	Eating	Greeting	Phoning	Photo	Posing	Purchases
LinKDE [14]	132.71	183.55	132.37	164.39	162.12	205.94	150.61	171.31
Li et al. [20]	-	134.13	97.37	122.33	-	166.15	-	-
Tekin *et al.* [36]	102.41	147.72	88.83	125.28	118.02	182.73	112.38	129.17
Zhou *et al.* [45]	87.36	109.31	87.05	103.16	116.18	143.32	106.88	99.78
Tekin *et al.* [35]	-	129.06	91.43	121.68	-	162.17	-	-
Ghezelghieh *et al.* [13]	80.30	80.39	78.13	89.72	-	-	-	-
Du *et al.* [10]	85.07	112.68	104.90	122.05	139.08	135.91	105.93	166.16
Park *et al.* [22]	100.34	116.19	89.96	116.49	115.34	149.55	117.57	106.94
Zhou *et al.* [44]	91.83	102.41	96.65	98.75	113.35	125.22	90.04	93.84
Ours	**67.38**	**71.95**	**66.70**	**69.07**	**71.95**	**76.97**	**65.03**	**68.30**
	Sitting	SittingDown	Smoking	Waiting	WalkDog	Walking	WalkTogether	Average
LinKDE [14]	151.57	243.03	162.14	170.69	177.13	96.60	127.88	162.14
Li et al. [20]	-	-	-	-	134.13	68.51	-	-
Tekin *et al.* [36]	138.89	224.9	118.42	138.75	126.29	**55.07**	65.76	124.97
Zhou *et al.* [45]	124.52	199.23	107.42	118.09	114.23	79.39	97.70	113.01
Tekin *et al.* [35]	-	-	-	-	130.53	65.75	-	-
Ghezelghieh *et al.* [13]	-	-	-	-	-	95.07	82.22	-
Du *et al.* [10]	117.49	226.94	120.02	117.65	137.36	99.26	106.54	126.47
Park *et al.* [22]	137.21	190.82	105.78	125.12	131.90	62.64	96.18	117.34
Zhou *et al.* [44]	132.16	158.97	106.91	94.41	126.04	79.02	98.96	107.26
Ours	**83.66**	**96.51**	**71.74**	**65.83**	**74.89**	59.11	**63.24**	**71.90**

Table 4: Quantitative comparison on Human3.6M. The numbers are the average 3D joint error (mm). Baseline numbers are taken from the respective papers. Note, several approaches use video for prediction rather than a single frame [36, 45, 10].

Figure 4: Sample qualitative results for all the datasets used in the empirical evaluation. First row: Human3.6M. Second row: HumanEva. Third row: KTH Football II. Fourth and fifth row: MPII. For each image, the original viewpoint and a novel viewpoint are shown. Red and green indicate left and right, respectively.

References

[1] A. Agarwal and B. Triggs. Recovering 3D human pose from monocular images. *PAMI*, 28(1):44–58, 2006. 1

[2] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 5

[3] M. Andriluka, S. Roth, and B. Schiele. Monocular 3D pose estimation and tracking by detection. In *CVPR*, 2010. 1

[4] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3D pictorial structures for multiple human pose estimation. In *CVPR*, 2014. 3

[5] L. Bo and C. Sminchisescu. Twin Gaussian processes for structured prediction. *IJCV*, 87(1-2):28–52, 2010. 7

[6] 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 *ECCV*, 2016. 1, 3, 7

[7] M. Burenius, J. Sullivan, and S. Carlsson. 3D pictorial structures for multiple view articulated pose estimation. In *CVPR*, 2013. 1, 3, 4, 5, 7

[8] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. In *CVPR*, 2016. 2, 4

[9] W. Chen, H. Wang, Y. Li, H. Su, D. Lischinsk, D. Cohen-Or, B. Chen, et al. Synthesizing training images for boosting human 3D pose estimation. In *3DV*, 2016. 3

[10] Y. Du, Y. Wong, Y. Liu, F. Han, Y. Gui, Z. Wang, M. Kankan-halli, and W. Geng. Marker-less 3D human motion capture with monocular image sequence and height-maps. In *ECCV*, 2016. 2, 7, 8

[11] A. Elhayek, E. de 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 *CVPR*, 2015. 3

[12] J. Gall, B. Rosenhahn, T. Brox, and H.-P. Seidel. Optimization and filtering for human motion capture. *IJCV*, 87(1):75–92, 2010. 3

[13] M. F. Ghezelghieh, R. Kasturi, and S. Sarkar. Learning camera viewpoint using CNN to improve 3D body pose estimation. In *3DV*, 2016. 2, 3, 8

[14] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *PAMI*, 36(7):1325–1339, 2014. 5, 8

[15] V. Kazemi, M. Burenius, H. Azizpour, and J. Sullivan. Multi-view body part recognition with random forests. In *BMVC*, 2013. 1, 3, 5

[16] I. Kostrikov and J. Gall. Depth sweep regression forests for estimating 3D human pose from images. In *BMVC*, 2014. 3, 5, 7

[17] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. In *AISTATS*, 2015. 4

[18] H.-J. Lee and Z. Chen. Determination of 3D human body postures from a single view. *CVGIP*, 30(2):148–168, 1985. 1

[19] S. Li and A. B. Chan. 3D human pose estimation from monocular images with deep convolutional neural network. In *ACCV*, 2014. 1, 2, 4, 6

[20] S. Li, W. Zhang, and A. B. Chan. Maximum-margin structured learning with deep networks for 3D human pose estimation. In *ICCV*, 2015. 1, 2, 5, 8

[21] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 2, 3, 4, 6

[22] S. Park, J. Hwang, and N. Kwak. 3D human pose estimation using convolutional neural networks with 2D pose information. In *ECCVW*, 2016. 2, 6, 8

[23] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Harvesting multiple views for marker-less 3D human pose annotations. *CVPR*, 2017. 3, 4

[24] T. Pfister, J. Charles, and A. Zisserman. Flowing convnets for human pose estimation in videos. In *ICCV*, 2015. 4

[25] I. Radwan, A. Dhall, and R. Goecke. Monocular image 3D human pose estimation under self-occlusion. In *ICCV*, 2013. 7

[26] V. Ramakrishna, T. Kanade, and Y. Sheikh. Reconstructing 3D human pose from 2D image landmarks. In *ECCV*, 2012. 1

[27] G. Rogez and C. Schmid. MoCap-guided data augmentation for 3D pose estimation in the wild. In *NIPS*, 2016. 1, 2, 3, 7

[28] M. Sanzari, V. Ntouskos, and F. Pirri. Bayesian image based 3D pose estimation. In *ECCV*, 2016. 7

[29] N. Sarafianos, B. Boteanu, B. Ionescu, and I. A. Kakadiaris. 3D human pose estimation: A review of the literature and analysis of covariates. *CVIU*, 152:1–20, 2016. 2

[30] L. Sigal, A. O. Balan, and M. J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87(1-2):4–27, 2010. 5

[31] L. Sigal, M. Isard, H. W. Haussecker, and M. J. Black. Loose-limbed people: Estimating 3D human pose and motion using non-parametric belief propagation. *IJCV*, 98(1):15–48, 2012. 3

[32] E. Simo-Serra, A. Quattoni, C. Torras, and F. Moreno-Noguer. A Joint Model for 2D and 3D Pose Estimation from a Single Image. In *CVPR*, 2013. 7

[33] E. Simo-Serra, A. Ramisa, G. Alenyà, C. Torras, and F. Moreno-Noguer. Single image 3D human pose estimation from noisy observations. In *CVPR*, 2012.

[34] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 4

[35] B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, and P. Fua. Structured prediction of 3D human pose with deep neural networks. In *BMVC*, 2016. 1, 2, 6, 8

[36] B. Tekin, A. Rozantsev, V. Lepetit, and P. Fua. Direct prediction of 3D body poses from motion compensated sequences. In *CVPR*, 2016. 2, 5, 7, 8

[37] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, 2014. 4

[38] R. Urtasun, D. J. Fleet, and P. Fua. 3D people tracking with gaussian process dynamical models. In *CVPR*, 2006. 1

[39] C. Wang, Y. Wang, Z. Lin, A. L. Yuille, and W. Gao. Robust estimation of 3D human poses from a single image. In *CVPR*, 2014. 7

[40] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016. 2, 4

[41] J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, and W. T. Freeman. Single image 3D interpreter network. In *ECCV*, 2016. 3, 4

[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 *CVPR*, 2016. 5, 7

[43] X. Zhou, S. Leonardos, X. Hu, and K. Daniilidis. 3D shape estimation from 2D landmarks: A convex relaxation approach. In *CVPR*, 2015. 1

[44] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei. Deep kinematic pose regression. In *ECCVW*, 2016. 2, 6, 8

[45] X. Zhou, M. Zhu, S. Leonardos, K. Derpanis, and K. Daniilidis. Sparseness meets deepness: 3D human pose estimation from monocular video. In *CVPR*, 2016. 1, 2, 5, 7, 8