TexturePose: Supervising Human Mesh Estimation with Texture Consistency

Georgios Pavlakos*, Nikos Kolotouros*, Kostas Daniilidis
University of Pennsylvania

Abstract

This work addresses the problem of model-based human pose estimation. Recent approaches have made significant progress towards regressing the parameters of parametric human body models directly from images. Because of the absence of images with 3D shape ground truth, relevant approaches rely on 2D annotations or sophisticated architecture designs. In this work, we advocate that there are more cues we can leverage, which are available for free in natural images, i.e., without getting more annotations, or modifying the network architecture. We propose a natural form of supervision, that capitalizes on the appearance constancy of a person among different frames (or viewpoints). This seemingly insignificant and often overlooked cue goes a long way for model-based pose estimation. The parametric model we employ allows us to compute a texture map for each frame. Assuming that the texture of the person does not change dramatically between frames, we can apply a novel texture consistency loss, which enforces that each point in the texture map has the same texture value across all frames. Since the texture is transferred in this common texture map space, no camera motion computation is necessary, or even an assumption of smoothness among frames. This makes our proposed supervision applicable in a variety of settings, ranging from monocular video, to multi-view images. We benchmark our approach against strong baselines that require the same or even more annotations that we do and we consistently outperform them. Simultaneously, we achieve state-of-the-art results among model-based pose estimation approaches in different benchmarks. The project website with videos, results, and code can be found at https://seas.upenn.edu/~pavlakos/projects/texturepose.

1. Introduction

In recent years, the area of human pose estimation has experienced significant successes for tasks with an increasing level of difficulty; 2D joint detection [30, 50], dense correspondence estimation [4] or even 3D skeleton reconstruction [26, 45]. Typically, as we ascend the pyramid of human understanding, we target more and more challenging tasks. As expected, the emergence of sophisticated parametric models of the human body, like SCAPE [6], SMPL(-X) [25, 32, 40], and Adam [17, 51], has really paved the way for full 3D pose and shape estimation from image data. And while this step has been well explored for video or multi-view data [13, 17], the ultimate goal is to reach the same level of analysis from a single image.

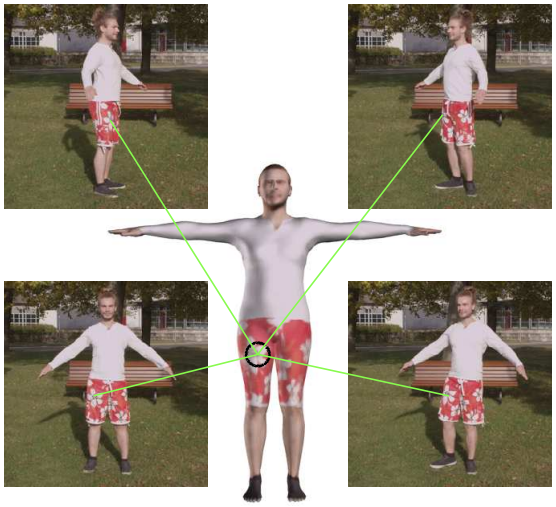

Figure 1: For a short video, or multi-view images of a person, a specific patch on the body surface has constant texture. This consistency can be formulated as an auxiliary loss in the training of a network for model-based pose estimation, and allows us to leverage information directly from raw pixels of natural images. Images and texture come from the People-Snapshot dataset [3].

Traditional optimization-based approaches, e.g., [8, 10, 23], have performed very reliably for model-based pose estimation. However, more recently, the interest has moved towards data-driven approaches regressing the parameters of the human body model, directly from images. Considering the lack of images with 3D shape ground truth for training, the main challenge is to identify reliable sources of supervision. Proposed methods [18, 31, 35, 46, 47, 53] have focused on leveraging all the available sources of 2D an-

* equal contribution

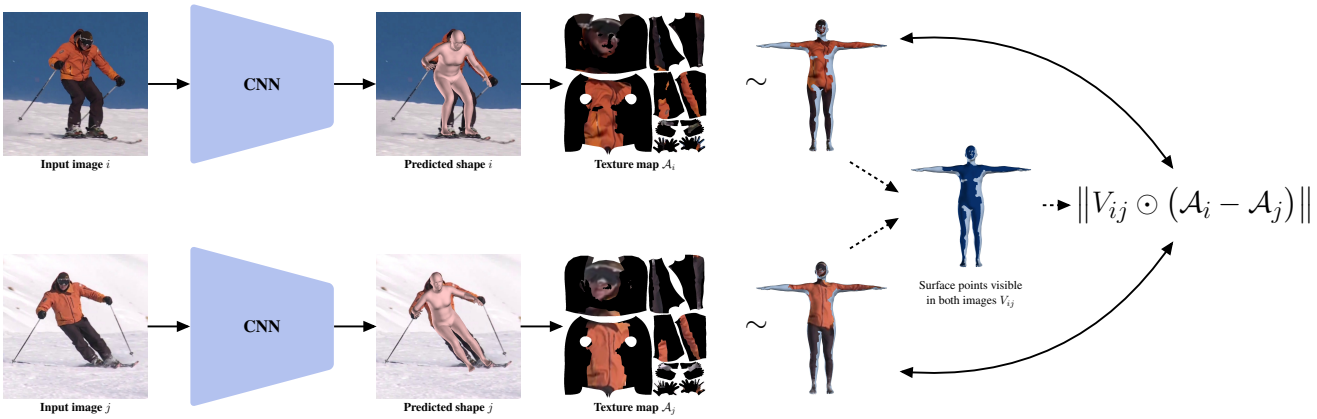

Figure 2: Overview of the proposed texture consistency supervision. Here, for simplicity, the input during training consists of two images i, j of the same person. The main assumption is that the appearance of the person does not change dramatically across the input images, (i.e., the frames come from a monocular video as in Figure 1, or from time-synchronized multi-view cameras). We apply our deep network on both images and estimate the shape of the person. Subsequently, we project the predicted shape on the image, and after inferring visibility for each point on the surface, we build the texture maps $\mathcal{A}_i$ and $\mathcal{A}_j$. The crucial observation, that the appearance of the person remains constant, translates to a texture consistency loss, forcing the two texture maps to be equal for all surface points V_{ij} that are visible in both images. This loss acts as supervision for the network and complements other weak losses that are typically used in the training.

notations like 2D keypoints, silhouettes, or semantic parts. Simultaneously, external sources of 3D data (e.g., MoCap and body scans) can also be useful, by applying learned priors [18], or decomposing the task in different architectural components [31, 35, 53]. In this work, instead of focusing on the available 2D annotations, or the appropriate way to employ external 3D data, the questions we ask are different. Can natural images alone provide us a useful cue for this task? Is there a form of supervision we can leverage without further annotations? Here, we argue, and demonstrate, that the answer to these questions is positive.

We present TexturePose, a way to leverage complementary supervision directly from natural images (Figure 2). The main observation is that the appearance of a person does not change significantly over small periods of time (e.g., during a short video). Our insight is that this appearance constancy enforces strong constraints in the estimated pose of each frame, which naturally translates to a powerful supervision signal that is useful for cases of monocular video or multi-view images. A critical component is the incorporation of a parametric model of the human body, SMPL [25], within our pipeline, allowing us to map the texture of the image to a generic texture map, which is independent of the shape and pose. Considering a network estimating the model parameters, during training, we generate the mesh and project it on the image. Through efficient computation, we are able to infer a (partial) texture map for each frame. Our novel supervision signal, based on texture consistency, enforces that the texture of each point of the texture map remains constant for all the frames of the same subject. This seemingly unimportant piece of information goes a long way and proves itself to be a crucial form of auxiliary supervision. We validate its importance in settings involving multiple views of the same subject, or monocular video with very weak annotations. In every case, we compare with approaches that have access to the same level of annotations (or potentially even more), and we consistently outperform them. Ultimately, this supervision allows us to outperform state-of-the-art approaches for model-based pose estimation from a single image.

Our contributions can be summarized as follows:

- We propose TexturePose, a novel approach to leverage complementary supervision from natural images through appearance constancy of each human across different frames.
- We demonstrate the effectiveness of our texture consistency supervision in cases of monocular video and multi-view capture, consistently outperforming approaches with access to the same or more annotations than we do.
- We achieve state-of-the-art results among model-based 3D human pose estimation approaches.

2. Related work

In this Section, we summarize the approaches that are more relevant to ours.

Model-based human pose estimation: Differently from skeleton-based 3D pose estimation, model-based human pose estimation involves a parametric model of the human body, e.g., SCAPE [6] or SMPL [25]. The goal

is to estimate the model parameters that give rise to a 3D shape which is consistent with image evidence. The initial works in this area [10, 43] as well some more recent approaches [8, 23, 52] were mainly optimization-based. Recently, the trend has shifted to directly regressing the model parameters from a single image using deep networks [18, 31, 35, 53]. Given the lack of images with 3D shape ground truth, these approaches typically rely on 2D annotations, like 2D keypoints, silhouettes and semantic parts, as well as external 3D data. Although, we believe there is great merit into using the bulk of already annotated data, in this paper we aspire to get beyond this data and explore complementary forms of supervision which are available also in unlabeled or weakly labeled data.

Multi-view pose estimation: Our goal in this work is not explicitly to estimate human pose from multiple views (in fact the work of Huang *et al*. [13] addressed this nicely in a model-based way). However, our approach is relevant to recent approaches leveraging multi-view consistency as a form of supervision to train deep networks. Pavlakos *et al*. [34] estimate 3D poses combining reliable 2D pose estimates, and treats them as pseudo ground truth to train a network for 3D human pose. Simon *et al*. [44] propose a similar approach to improve a hand keypoint detector given multi-view data. Rhodin *et al*. [39] learn 3D pose estimation by enforcing the pose consistency in all views. On the other hand, follow-up work from Rhodin *et al*. [38], uses multiple views to learn a representation of 3D human pose in an unsupervised manner. In contrast to the above works, we believe that our approach offers much greater opportunities to leverage multi-view consistency. The incorporation of a parametric model allows us to go beyond body joint consistency, by leveraging shape and texture consistency. Simultaneously, instead of learning a new representation from multi-view data, we choose to maintain the SMPL representation, and only leverage the collective power of data to better regress the parameters of this representation.

Supervision signals: While we have already discussed some aspects of the supervision typically employed for 3D human pose estimation, here we attempt to extend the discussion particularly to the varying levels of supervision used by different works. Full body pose and shape supervision is typically available only in synthetic images [48], or images with successful body fits [23]. Weak supervision provided by 2D annotations is typical, with different works employing 2D keypoints, silhouettes and semantic parts [18, 31, 35, 46]. Non-parametric approaches typically use extra supervision from 2D keypoint annotation [12, 45, 55], while some recent works leverage ordinal depth relations of the joints [33, 41]. Multi-view consistency is also well explored as discussed earlier [21, 34, 38, 39, 44]. In terms of pose priors, Zhou *et al*. [55] use weak symmetry constraints, while Kanazawa *et al*. [18] incorporates a learning-based prior on pose and shape parameters using adversarial networks. In contrast to the above, instead of using additional annotations or exploiting external information, our goal is to leverage all the information that is available in natural images. This of course does not exclude the use of other supervision forms. In fact, we demonstrate that our approach can properly complement typical supervision signals (e.g., 2D keypoints, pose priors), and improve performance only by additionally enforcing texture consistency.

Texture-based approaches: The idea of using texture to guide pose estimation goes back at least to the work of Sidenbladh *et al*. [42], where texture consistency was used for tracking. More recently, Bogo *et al*. [9] use high resolution texture information to improve registration alignments. Guo *et al*. [11] also enforce photometric consistency to recover accurate human geometry over time. Alldieck *et al*. [1, 2, 3] focus on estimating the texture for human models. In the work of Kanazawa *et al*. [19], texture is employed to learn a parametric model of bird shapes. While we share similar intuitions with the above works, here we propose to use texture as a supervisory signal to guide and improve learning for 3D human pose and shape estimation.

Finally, to put our work in a greater context, the idea of appearance constancy is popular also beyond human pose estimation, e.g., in approaches for unsupervised learning of depth, ego-motion and optical flow [15, 28, 36, 54]. A key difference is that while they estimate the structure of the world in a non-parametric form (depth map), we instead inject some domain knowledge (i.e., assuming a human pose estimation task) and we leverage a model, SMPL, that helps us explain the image observations. A similar motivation is shared with the work of Tung *et al*. [47]. However, our approach is more flexible, since they require keypoints as input to their network, frames should be continuous to allow for motion extraction, while they eventually rely on a separate network for optical flow computation. Simultaneously, we present a more generic framework, which can be applied for monocular video or multi-view images alike.

3. Technical approach

In this Section, we start with a short introduction about the representation we use and the basic notation (Subsection 3.1). Then, we describe the regression architecture (Subsection 3.2). We continue with the formulation of texture consistency, and the corresponding loss (Subsection 3.3). Next, we describe the additional losses we can incorporate when we process images from monocular or multi-view input (Subsection 3.4). Finally, we provide an overview of the complete pipeline (Subsection 3.5), and discuss potential weaknesses of our approach (Subsection 3.6).

3.1. Representation

SMPL: The SMPL model [25] is a parametric model of the human body. Given the input parameters for pose $\boldsymbol{\theta}$, and shape $\boldsymbol{\beta}$, the model defines a function $\mathcal{M}(\boldsymbol{\theta}, \boldsymbol{\beta})$ which outputs the body mesh $M \in \mathbb{R}^{3\times N}$, with $N = 6890$ vertices. The body joints X are expressed as a linear combination of the mesh vertices, so using a pre-trained linear regressor W, we can map from the mesh to k joints of interest $X \in \mathbb{R}^{3\times k} = WM$.

Texture map: The meshes produced by SMPL are deformations of an original template T. A corresponding UV map un-warps the template surface onto an image, $\mathcal{A}$, which is the texture map. Each pixel t of this texture map is also called texel. By construction, the mapping between texels and mesh surface coordinates is fixed and independent of changes in 3D surface geometry.

Camera: The camera we use follows the weak perspective camera model. The parameters of interest are denoted with $\boldsymbol{\pi}$ and include the global orientation $R \in \mathbb{R}^{3\times 3}$, scale $s \in \mathbb{R}$, and translation $t \in \mathbb{R}^2$. Given these parameters, the 2D projection x of the 3D joints X is expressed as:

$$x = \boldsymbol{\pi}(X) = s\Pi(RX) + t, \tag{1}$$

where Π stands for the orthographic projection.

3.2. Regression model

Our goal is to learn a predictor f, here realized by a deep network, that given a single image I, it maps it to the pose and shape parameters of the person on the image. More concretely, the output of the network consists of a set of parameters $\boldsymbol{\Theta} = f(I)$, where $\boldsymbol{\Theta} = [\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\pi}]$. Here, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ indicate the SMPL pose and shape parameters, and $\boldsymbol{\pi}$ are the camera parameters. Our deep network follows the architecture of Kanazawa *et al*. [18], with the exception of the output, which in our case regresses 3D rotations using the representation proposed by Zhou *et al*. [56].

3.3. TexturePose

Given $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ we can generate a mesh M and the corresponding 3D joints X. The mesh can be projected to the image using the estimated camera parameters $\boldsymbol{\pi}$. Through efficient computation [29], we can infer the visibility for each point on the surface, and as a result, for every texel t of the texture map $\mathcal{A}$. Let us denote with v_t the inferred visibility of texel t on the texture map $\mathcal{A}$. The collection of all visibility indices v_t can be arranged in a binary mask V, the visibility mask. Considering each point P_t on the mesh surface, we can estimate its image projection using the camera parameters, $p_t = \boldsymbol{\pi}(P_t)$. For the visible points, we can estimate their texture via bilinear sampling from the image I, so $a_t = G(I; p_t)$, where G is a bilinear sampling kernel. The collection of all values a_t, i.e., texture values for every texel t, constitute the texture map $\mathcal{A}$.

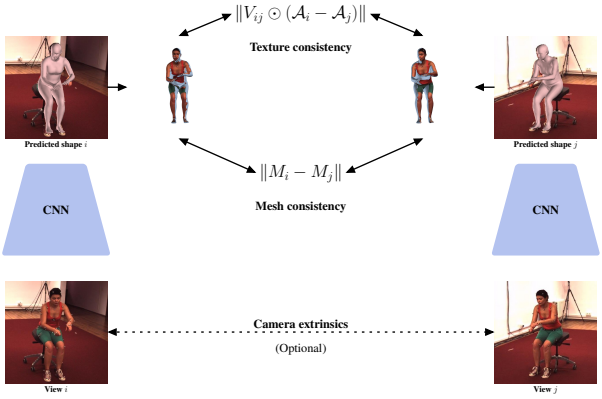

Figure 3: With our formulation, training with images from a multi-camera system is similar to training with images from monocular video (Figure 2). The main additional consistency constraint is that the subject has the same 3D shape (same body mesh), which means that we can apply a per-vertex loss between the two mesh predictions. Before applying the predicted global orientation, the mesh predictions are in the same canonical orientation, so we can apply our loss directly on the mesh predictions. In case the extrinsics are provided, we can transform the second mesh to the frame of the first view, and then apply the same loss.

Let us now assume that we have access to two images i, j of the same person. Using the procedure above, we can estimate the two texture maps $\mathcal{A}_i, \mathcal{A}_j$, along with the corresponding visibility masks V_i, V_j. Let us denote with $V_{ij} = V_i \odot V_j$ the mask of the surface points that are visible in both views. Then the texture consistency loss can be simply defined as:

$$L_{\text{texture cons}} = ||V_{ij} \odot (\mathcal{A}_i - \mathcal{A}_j)||. \tag{2}$$

This loss enforces that the texture should be the same for texels (or equivalently, points on the surface) that are visible in both images. Since visibility masks are used only to mask-out the texels that should not contribute to the loss, visibility computation does not have to be differentiable.

3.4. Beyond texture

Monocular: In the monocular case, the texture consistency loss is applied between pairs of frames for the same subject. Beyond the texture consistency, we can also enforce that the shape parameters of the subject remain the same for all pairs of frames. This shape consistency can be enforced with the following loss function:

$$L_{\text{shape cons}} = ||\boldsymbol{\beta}_i - \boldsymbol{\beta}_j||. \tag{3}$$

Furthermore, we want to guarantee that we get a valid 3D shape, i.e., the estimated pose and shape parameters of the parametric model lie in the space of valid poses and shapes respectively. To enforce this, we use the adversarial prior

of Kanazawa *et al*. [18], which factorizes the model parameters into: (i) pose parameters $\boldsymbol{\theta}$, (ii) shape parameters $\boldsymbol{\beta}$, and (iii) per-part relative rotations $\boldsymbol{\theta}_i$, that is one 3D rotation for each of the 23 joints of SMPL. In the end, we train a discriminator D_k for each factor of the body model. The generator loss can then be expressed as:

$$L_{\text{adv prior}} = \sum_k (D_k(\boldsymbol{\Theta}) - 1)^2. \quad (4)$$

Depending on the availability of additional 2D keypoint annotations, we can also enforce that the 3D joints project close to the annotated 2D keypoints. We get the projection of the 3D joints X to the 2D locations x, based on Eq. 1. Then, the 2D-3D consistency can be expressed as:

$$L_{\text{2D}} = ||x - x_{\text{gt}}||, \quad (5)$$

where x_{gt} are the ground truth 2D joints. Finally, adding smoothness on the pose parameters is also possible, but we avoid it, to keep our approach more generic and applicable even in settings where the frames are not consecutive.

Multiple views: When we have access to multiple views i and j of a subject at the same time instance, then all the above losses remain relevant. The main additional constraint we need to enforce is that the pose of the person is the same across all viewpoints. This could be incorporated, by simply forcing all the pose parameters to have the same value. In contrast to that, we observed that a loss applied directly on the mesh vertices behaves much better (Figure 3). This can be formulated as a simple per-vertex loss:

$$L_{\text{mesh cons}} = ||M_i - M_j||. \quad (6)$$

Remember that M_i, M_j do not include the global orientation estimates R_i, R_j, so both meshes are in the canonical orientation, meaning that we can compare them directly. This loss effectively reflects the more generic case, where no knowledge of the camera extrinsics is available for the multi-view system. If extrinsic calibration is also known, then we simply need to apply the global pose estimates R_i, R_j, transform the second mesh to the coordinate system of the first mesh and then use the same per-vertex loss.

3.5. Complete pipeline

Our network is trained using batches of images. When we want to use a short sequence in training, or a few time-synchronized viewpoints, we include all the frames of interest in the batch. Typically, for monocular video, we include five consecutive frames, while for multi-view images, we use as many viewpoints are available at a specific time instance (typically four for Human3.6M). Conveniently, during testing, we can process each frame independently, without the need for video or multi-view input.

Depending on the setting, and making sure that we are compatible with prior work, we can also augment our batches with images that have stronger supervision (e.g., full 3D pose is known). Since the texture consistency assumption alone keeps the problem pretty underconstrained, similar to prior works (e.g., [39, 38]), we found that it was useful to have stronger supervision in at least a few examples. For fair comparisons, in the empirical evaluation, we make sure that we use the same, or strictly less annotations than what prior work is using.

3.6. Shortcomings

Although we empirically demonstrate the significant value of TexturePose (Section 4), it is fair to also identify some of the shortcomings of our approach. For example the constant appearance assumption can easily be violated (e.g., due to illumination or viewpoint changes). Moreover, motion blur is common and can also decrease the level of "clean" pixels we can benefit from. Finally, our approach makes an assumption that no object occludes the person. Since we do not account for the potential occlusions, we can easily fill the texture map with the texture of the occluding object. Although occlusions are not very typical in most of the images for the datasets we use, this can be a source of potential error given a new video for training. The work of Ranjan *et al*. [37] addresses a similar problem in the context of Structure from Motion, and we believe that a similar approach should be applicable in our setting as well.

4. Empirical evaluation

In this Section, we summarize the empirical evaluation of our approach. First, we provide more details about the datasets we employ for training and evaluation (Subsection 4.1), and then we present quantitative (Subsection 4.2) and qualitative results (Subsection 4.3).

4.1. Datasets

For the majority of our ablation studies, we used the Human3.6M dataset [14]. Additionally, we used training data from the MPII 2D human pose dataset [5], while LSP dataset [16] was employed only to evaluate our approach. In the Sup.Mat. we present more extensive experiments leveraging the recently introduced VLOG-People and InstaVariety datasets [20] for training, as well as the 3DPW dataset [49] for evaluation.

Human3.6M: It is an indoor benchmark for 3D human pose estimation. It includes multiple subjects performing daily actions like Eating, Smoking and Walking. It provides videos from four calibrated, time-synchronized cameras, making it easy to evaluate the different aspects of our approach both in the monocular and the multi-view setting. For training, we used subjects S1, S5, S6, S7 and S8, unless otherwise stated. Being consistent with prior work [18], the

evaluation is done on subjects S9 and S11, considering Protocol 1 [38, 39] and Protocol 2 [18].

MPII: It is an in-the-wild dataset for 2D human pose estimation, providing only the 2D joint locations for each person. Previous works [18, 35] typically employ this dataset because of the large number of 2D keypoint annotations. One typically unexplored advantage of this dataset is the fact that it also provides the neighboring frames of the video that includes the annotated frame. We see this as a large pool of unlabeled data, that we can leverage for free, and we demonstrate their effectiveness in training our models. We call this set "MPII video" and consists of the annotated frames for each video, along with four more frames (two before, two after), which come with no labels.

LSP: It is also an in-the-wild dataset for 2D human pose estimation, but of much smaller scale compared to MPII. We employ LSP only for evaluation, where we make use of its test set. Particularly, given our shape prediction, we project it back to the image and we evaluate silhouette and part segmentation accuracy. For this evaluation, we use the segmentation labels provided by [23].

4.2. Quantitative evaluation

Ablative studies: We start with Human3.6M where we initially treat all the images as frames of monocular sequences. One strong baseline, inspired from the "unpaired" setting of [18], assumes that the network has access to the 2D joints for each image, and an independent dataset of 3D poses, but no image with corresponding 3D ground truth is available. We train the network with 2D reprojection loss, while we also enforce an adversarial prior for pose and shape parameters such that the predicted poses/shapes are close to the poses/shapes in the dataset. As we can see, in Table 1 (first row), this gives us decent performance. If we also apply our texture consistency loss, over the frames of the short clips, then we get significant improvement (second row). Finally, to put these results into context, we also train the same architecture providing full 3D pose and shape ground truth for each image (third row). As expected, this ideal version is performing better, but our texture consistency loss managed to close the gap between the weakly- and the fully-supervised setting.

A similar experiment attempts to investigate the effect of leveraging texture consistency, but this time from in-the-wild videos. To this end, we use the frames of MPII video applying our texture consistency. The results for our experiments are presented in Table 2. The initial baseline (first row) is the same as in Table 1, and uses full 3D ground truth from Human3.6M for training. The next thing we want to investigate is whether adding purely unlabeled video can improve performance. So, for the next baseline (second row), we provide no labels for the in-the-wild frames, but enforce texture consistency. Interestingly, the model does get improved just by seeing more unlabeled data simply by enforcing texture consistency. Unfortunately, we observed that although the performance improves for Human3.6M, when we apply the same model to in-the-wild images, it achieves mediocre results qualitatively. We believe that at least a few labels should be necessary to make the model generalize better. To this end, we conduct two more experiments, adding annotations for one frame of the MPII video sequences. In the first experiment (third row), we add the annotation for the frame, but no texture consistency loss is enforced, while for the second one (fourth row), we both add the annotation for the frame, and we activate the texture consistency loss. As we can see, adding the unlabeled frames helps by default when combined with a texture consistency loss, and gives a solid performance improvement, making the model appropriate both for Human3.6M and for in-the-wild images, as we will present later.

	P1	P2
2D keypoints + GAN prior	93.0	79.1
2D keypoints + GAN prior + Texture Consistency	80.2	76.2
3D Ground Truth	64.8	63.9

Table 1: The effect of texture consistency for monocular input on Human3.6M (Protocols 1 & 2). The numbers are mean reconstruction errors in mm. Using only 2D annotations on each frame and an adversarial pose/shape prior, we get reasonable performance. Simply providing more video frames instead of single frames *without any additional annotation*, we are able to get an important performance improvement, because of the texture consistency loss. As a lower limit, we present results when the ground truth 3D pose and shape parameters are available for each image during training. Although this last version uses explicitly stronger annotations, we are able to shrink the gap between the baselines that train with 2D and 3D annotations respectively.

The same findings extend also to the case that we add more video data that only contain automatic pseudo-annotations [20, 7]. We present our results using two recently published datasets for training, i.e., VLOG-People and InstaVariety [20] in the Sup.Mat., along with the 3DPW dataset [49] for additional evaluation.

Comparison with the state-of-the-art: For the comparison with the state-of-the-art, we use our best model from the previous experiment (last row of Table 2). The results are presented in Table 3. Our method outperform the previous baselines. Of course, we use MPII video for training which is not used by the other approaches, but making it possible to leverage the unlabeled frames is possible due to the texture consistency loss, which is one of our contributions. At the same time, other approaches also employ explicitly more annotations than we do, e.g., [18] has access to more images with 2D annotations (COCO [24]) and 3D annotations (MPI-INF-3DHP [27]), which we do not use.

		P1	P2
H36M		64.8	63.9
H36M	+ MPII videos (+texture)	60.1	58.6
H36M	+ MPII 2D	54.1	51.6
H36M	+ MPII videos (+texture) + MPII 2D	51.3	49.7

Table 2: Evaluation on the Human3.6M dataset (Protocols 1 & 2), indicating the effect of TexturePose, when we incorporate in-the-wild videos (MPII) in our training. Adding unlabeled video frames and enforcing texture consistency (row 2) improves Human3.6M evaluation, but the qualitative performance for in-the-wild images is mediocre. If we add a sparse set of 2D keypoint annotations (row 3), the performance can improve. However, the most interesting aspect is that by simply adding the sparse 2D keypoints labels, the *unlabeled* video frames *and* enforcing texture consistency (row 4), we can improve performance even more, meaning that extra unlabeled data can always be helpful.

	Rec. Error
Lassner *et al.* [23]	93.9
Pavlakos *et al.* [35]	75.9
NBF [31]	59.9
HMR [18]	56.8
Kanazawa *et al.* [20]	56.9
Arnab *et al.* [7]	54.3
Kolotouros *et al.* [22]	50.1
Ours	**49.7**

Table 3: Evaluation on the Human3.6M dataset (Protocol 2). The numbers are mean reconstruction errors in mm. We compare with regression approaches that output a mesh of the human body. Our approach achieves state-of-the-art results.

Moreover, we evaluate our approach on the LSP dataset using the same model (which has never been trained with images from LSP). Although in-the-wild, this dataset gives us access to segmentation annotations, so that we can evaluate shape estimation implicitly through mesh reprojection. The complete results are presented in Table 4. Here, we outperform the regression-based baseline of [18] which is more relevant to ours and we are also very competitive to the optimization-based approaches, which explicitly optimize for the image-model alignment, so they tend to perform better under these metrics.

Finally, we also compare with baselines trained with multiple-views. We follow the Protocol of Rhodin *et al.* [38, 39], training with full 3D supervision for S1 and employ the other training users without any further annotations, other than extrinsic calibration. The results are presented in Table 5. We successfully outperform both baselines. It is interesting that both [38, 39] are non-parametric approaches, and we are still able to outperform them, considering that strong non-parametric baselines [26, 33] typically perform better than parametric approaches [18, 35] (at least under the 3D joints metrics). We believe that this is exactly because we are able to leverage cues that are not an option for 3D skeleton baselines, e.g., they cannot map texture to a skeleton figure, as we can do with a mesh surface.

	FB Seg.		Part Seg.	
	acc.	f1	acc.	f1
SMPLify *oracle* [8]	92.17	0.88	88.82	0.67
SMPLify [8]	91.89	0.88	87.71	0.64
SMPLify on [35]	92.17	0.88	88.24	0.64
HMR [18]	91.67	0.87	87.12	0.60
Ours	91.82	0.87	89.00	0.67

Table 4: Evaluation on foreground-background and six-part segmentation on the LSP test set. The numbers are accuracies and f1 scores. Our approach outperforms the strong regression-based baseline of [18] across the Table, and it is very competitive to the optimization baselines based on SMPLify (which typically have advantage for tasks involving image-model alignment like this). The numbers for the first two rows are taken from [23].

	MPJPE	NMPJPE	Rec. Error
Rhodin *et al.* [39]	n/a	153.3	128.6
Rhodin *et al.* [38]	131.7	122.6	98.2
Ours	**110.7**	**97.6**	**74.5**

Table 5: Evaluation on Human3.6M (Protocol 1) for methods trained on multiple views. The numbers are mm in various metrics. We follow the protocol of [39, 39], using full 3D ground truth for S1, and leveraging the other subjects as unlabeled data, where only the camera calibration is known.

4.3. Qualitative evaluation

A variety of qualitative results of our results are provided in Figure 4 as well as in the Sup.Mat.

5. Summary

In this paper, we presented TexturePose, an approach to train neural networks for model-based human pose estimation by leveraging supervision directly from natural images. Effectively, we capitalize on the observation that the appearance of a person does not change dramatically within a short video or for images from multiple views. This allows us to apply a texture consistency loss which acts as a form of auxiliary supervision. This generic formulation makes our approach particularly flexible and applicable in monocular video and multi-view images alike. We compare TexturePose with different baselines requiring the same (or

Figure 4: Qualitative Results. Rows 1-5: LSP dataset. Rows 6-7: H36M dataset

larger) amount of annotations and we consistently outperform them, achieving state-of-the-art results across model-based pose estimation approaches. Going forwards, we believe that these weak supervision signals could really help us scale our training by leveraging weakly annotated or purely unlabeled data. Having already identified the shortcomings of our approach (Subsection 3.6), it is a great challenge to identify ways to go beyond them.

Acknowledgements: We gratefully appreciate support through the following grants: NSF-IIP-1439681 (I/UCRC), NSF-IIS-1703319, NSF MRI 1626008, ARL RCTA W911NF-10-2-0016, ONR N00014-17-1-2093, ARL DCIST CRA W911NF-17-2-0181, the DARPA-SRC C-BRIC, by Honda Research Institute and a Google Daydream Research Award.

References

[1] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single RGB camera. In *CVPR*, 2019. 3

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *3DV*, 2018. 3

[3] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3D people models. In *CVPR*, 2018. 1, 3

[4] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018. 1

[5] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 5

[6] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: shape completion and animation of people. *ACM transactions on graphics (TOG)*, 24(3):408–416, 2005. 1, 2

[7] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3D human pose estimation in the wild. In *CVPR*, 2019. 6, 7

[8] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*, 2016. 1, 3, 7

[9] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J Black. Dynamic FAUST: Registering human bodies in motion. In *CVPR*, 2017. 3

[10] Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. Estimating human shape and pose from a single image. In *CVPR*, 2009. 1, 3

[11] Kaiwen Guo, Feng Xu, Tao Yu, Xiaoyang Liu, Qionghai Dai, and Yebin Liu. Real-time geometry, albedo, and motion reconstruction using a single rgb-d camera. *ACM Transactions on Graphics (TOG)*, 36(3):32, 2017. 3

[12] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Gerard Pons-Moll, and Christian Theobalt. In the wild human pose estimation using explicit 2D features and intermediate 3D representations. In *CVPR*, 2019. 3

[13] Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V. Gehler, Javier Romero, Ijaz Akhter, and Michael J. Black. Towards accurate marker-less human shape and pose estimation over time. In *3DV*, 2017. 1, 3

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

[15] J Yu Jason, Adam W Harley, and Konstantinos G Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In *ECCVW*, 2016. 3

[16] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. 5

[17] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *CVPR*, 2018. 1

[18] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 1, 2, 3, 4, 5, 6, 7

[19] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 3

[20] Angjoo Kanazawa, Jason Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. In *CVPR*, 2019. 5, 6, 7

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

[22] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019. 7

[23] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3D and 2D human representations. In *CVPR*, 2017. 1, 3, 6, 7

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 6

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):248, 2015. 1, 2, 4

[26] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3D human pose estimation. In *ICCV*, 2017. 1, 7

[27] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved cnn supervision. In *3DV*, 2017. 6

[28] Simon Meister, Junhwa Hur, and Stefan Roth. UnFlow: Unsupervised learning of optical flow with a bidirectional census loss. In *AAAI*, 2018. 3

[29] MPI-IS. Mesh processing library. https://github.com/MPI-IS/mesh. 4

[30] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 1

[31] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *3DV*, 2018. 1, 2, 3, 7

[32] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, 2019. 1

[33] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3D human pose estimation. In *CVPR*, 2018. 3, 7

[34] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Harvesting multiple views for marker-less 3D human pose annotations. In *CVPR*, 2017. 3

[35] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *CVPR*, 2018. 1, 2, 3, 6, 7

[36] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *CVPR*, 2019. 3

[37] Anurag Ranjan, Javier Romero, and Michael J Black. Learning human optical flow. In *BMVC*, 2018. 5

[38] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3D human pose estimation. In *ECCV*, 2018. 3, 5, 6, 7

[39] Helge Rhodin, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua. Learning monocular 3D human pose estimation from multi-view images. In *CVPR*, 2018. 3, 5, 6, 7

[40] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 36(6):245, 2017. 1

[41] Matteo Ruggero Ronchi, Oisin Mac Aodha, Robert Eng, and Pietro Perona. It's all relative: Monocular 3D human pose estimation from weakly supervised data. In *BMVC*, 2018. 3

[42] Hedvig Sidenbladh, Michael J Black, and David J Fleet. Stochastic tracking of 3D human figures using 2D image motion. In *ECCV*, 2000. 3

[43] Leonid Sigal, Alexandru Balan, and Michael J Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *NIPS*, 2008. 3

[44] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017. 3

[45] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018. 1, 3

[46] Vince Tan, Ignas Budvytis, and Roberto Cipolla. Indirect deep structured learning for 3D human body shape and pose prediction. In *BMVC*, 2017. 1, 3

[47] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *NIPS*, 2017. 1, 3

[48] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 3

[49] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 5, 6

[50] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016. 1

[51] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *CVPR*, 2019. 1

[52] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3D pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *CVPR*, 2018. 3

[53] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu. Deep network for the integrated 3D sensing of multiple people in natural images. In *NIPS*, 2018. 1, 2, 3

[54] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 3

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

[56] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 4